Evolutionary analysis of a complex multigene family

by semi-automated mining of public sequence databases

Denis BAURAIN

Eukaryotic Phylogenomics
Dept of Life Sciences – ULg

Lultzhausen (GDL) – June 19-20, 2012

Acknowledgments

Plant Physiol. 2012 Feb;158(2):546-60. Epub 2011 Dec 12.

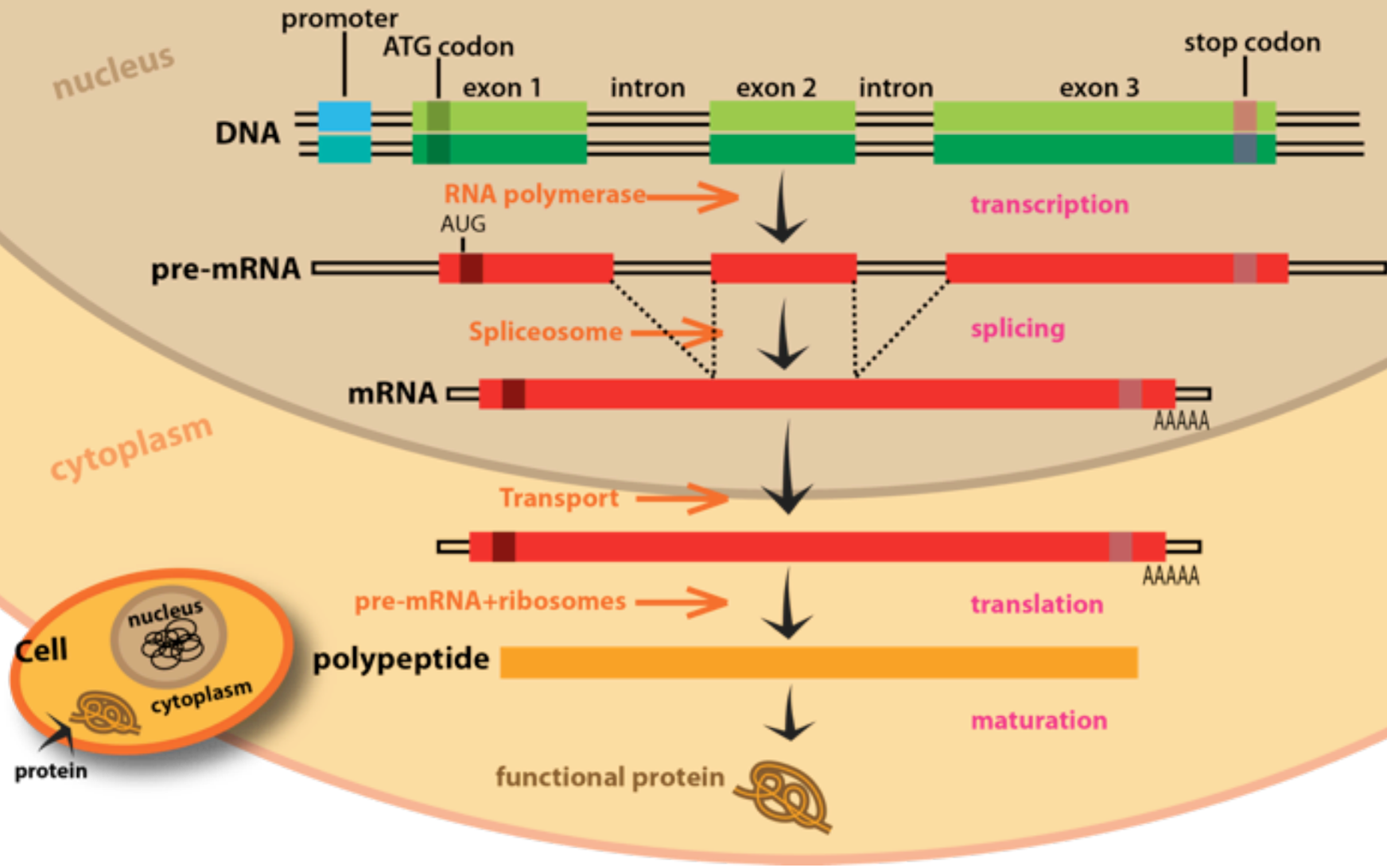
A single ancient origin for prototypical serine/arginine-rich splicing factors.

Califice S, Baurain D, Hanikenne M, Motte P.

Laboratory of Functional Genomics and Plant Molecular Imaging and Centre for Assistance in Technology of Microscopy, Department of Life Sciences, Institute of Botany, University of Liège, B-4000 Liege, Belgium.

Abstract

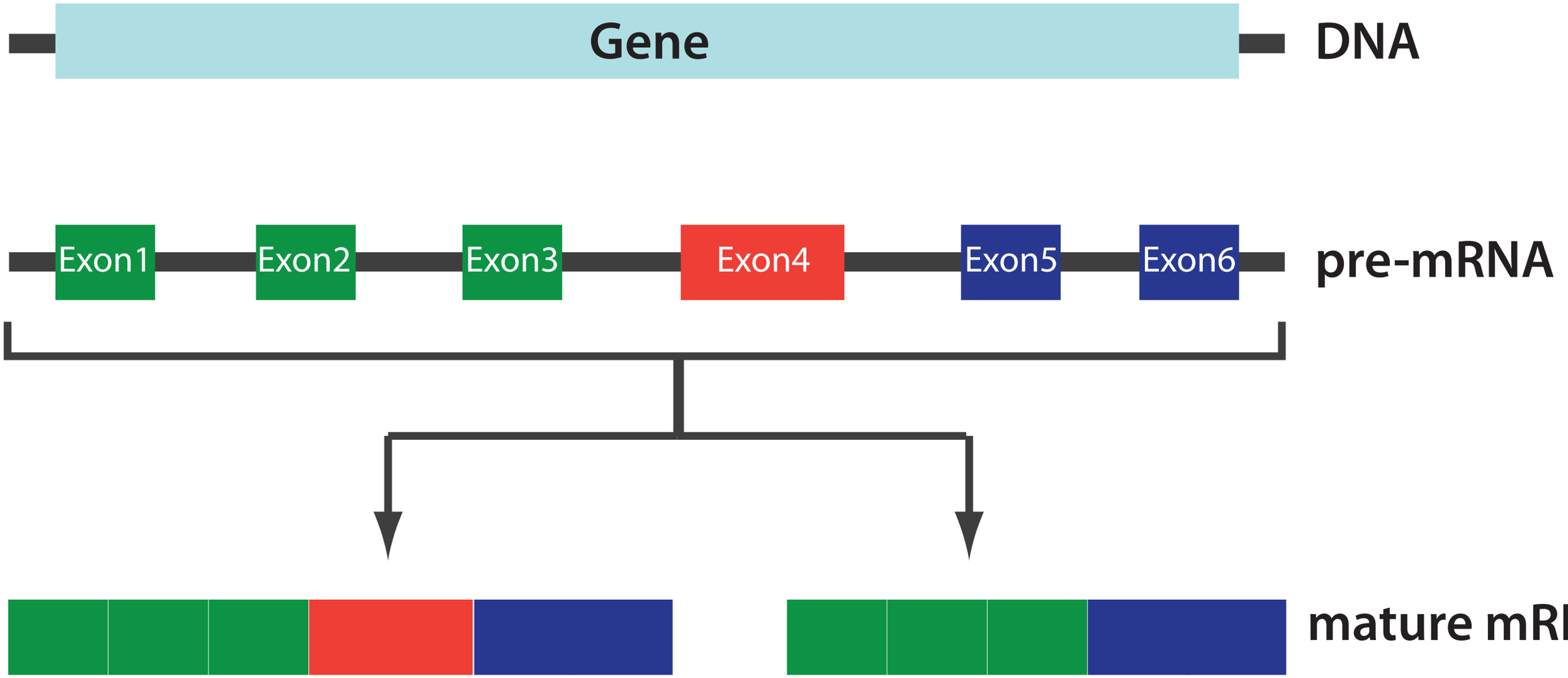
Eukaryotic precursor mRNA splicing is a process involving a very complex RNA-protein edifice. Serine/arginine-rich (SR) proteins play essential roles in precursor mRNA constitutive and alternative splicing and have been suggested to be crucial in plant-specific forms of developmental regulation and environmental adaptation. Despite their functional importance, little is known about their origin and evolutionary history. SR splicing factors have a modular organization featuring at least one RNA recognition motif (RRM) domain and a carboxyl-terminal region enriched in serine/arginine dipeptides. To investigate the evolution of SR proteins, we infer phylogenies for more than 12,000 RRM domains representing more than 200 broadly sampled organisms. Our analyses reveal that the RRM domain is not restricted to eukaryotes and that all prototypical SR proteins share a single ancient origin, including the plant-specific SR45 protein. Based on these findings, we propose a scenario for their diversification into four natural families, each corresponding to a main SR architecture, and a dozen subfamilies, of which we profile both sequence conservation and composition. Finally, using operational criteria for computational discovery and classification, we catalog SR proteins in 20 model organisms, with a focus on green algae and land plants. Altogether, our study confirms the homogeneity and antiquity of SR splicing factors while establishing robust phylogenetic relationships between animal and plant proteins, which should enable functional analyses of lesser characterized SR family members, especially in green plants.



A eukaryotic gene is arranged in a patchwork of coding (exons) and non-coding sequences (introns).

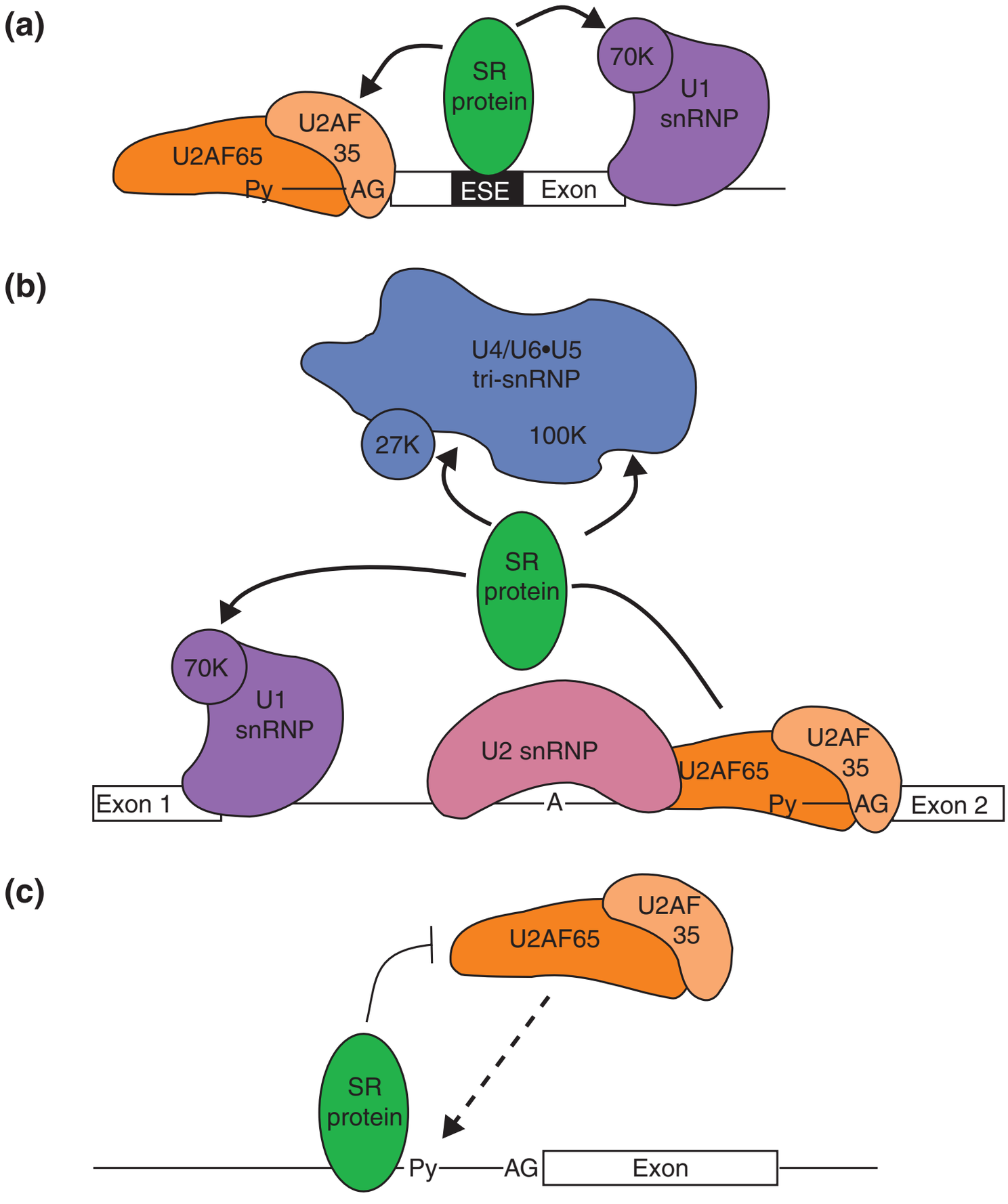
Introns are eliminated while exons are spliced together to yield the mature mRNA used for protein synthesis.

http://reflexions.ulg.ac.be/



In many organisms, almost all genes can lead to different proteins through a process termed alternative splicing.

http://reflexions.ulg.ac.be/



Splicing is carried out by a dynamic macromolecular complex known as the spliceosome.

Ser/Arg-rich (SR) splicing factors are involved in both constitutive and alternative splicing.

Shepard and Hertel (2009) *Genome Biol* 10:242

>Homo_sapiens@ENSP00000293233 [SC35]
MSYGRPPPDVEGMTSLKVDNLTYRTSPDTLRRVFEKY
GRVGDVYIPRDRYTKESRGFAFVRFHDKRDAEDAMDA
MDGAVLDGRELRVQMARYGRPPDSHHSRRGPPPRRYG
GGGYGRRSRSPRRRRRSRSRSRSRSRSRSRSRSRYSRSK
SRSRTRSRSRSTSKSRSARRSKSKSSSVSRSRSRSRS
RSRSRSPPPVSKRESKSRSRSKSPPKSPEEEGAVSS

All SR proteins have at least one N-term RNA Recognition Motif (RRM) and one C-term RS domain of low complexity enriched in Arg/Ser repeats.

>Homo_sapiens@ENSP00000325905 [9G8]
MSRYGRYGGETKVYVGNLGTGAGKGELERAFSYYGPL
RTVWIARNPPGFAFVEFEDPRDAEDAVRGLDGKVICG
SRVRVELSTGMPRRSRFDRPPARRPFDPNDRCYECGE
KGHYAYDCHRYSRRRRSRSRSRSHSRSRGRRYSRSRS
RSRGRRSRSASPRRSRSISLRRSRSASLRRSRSGSIK
GSRYFQSPSRSRSRSRSISRPRSSRSKSRSPSPKRSR
SPSGSPRRSASPERMD

But some SR splicing factors feature other domains, such as the CCHC zinc-knuckle (ZnK) motif.

ASF/SF2 type

SRp32, SRp33a, SRp33b, (SR20)

SC35 type

SC35a, SC35b, SC35c,
SCL25, SCL26, SCL30a, SCL30b

9G8 type

RSZp21a, RSZp21b, RSZp23

ZnK

Two RRMs (Plant specific)

RSp29, RSp33

Two Zinc Knuckles (Plant specific)

RSZ36, RSZ37a, RSZ37b, RSZ39

2x ZnK

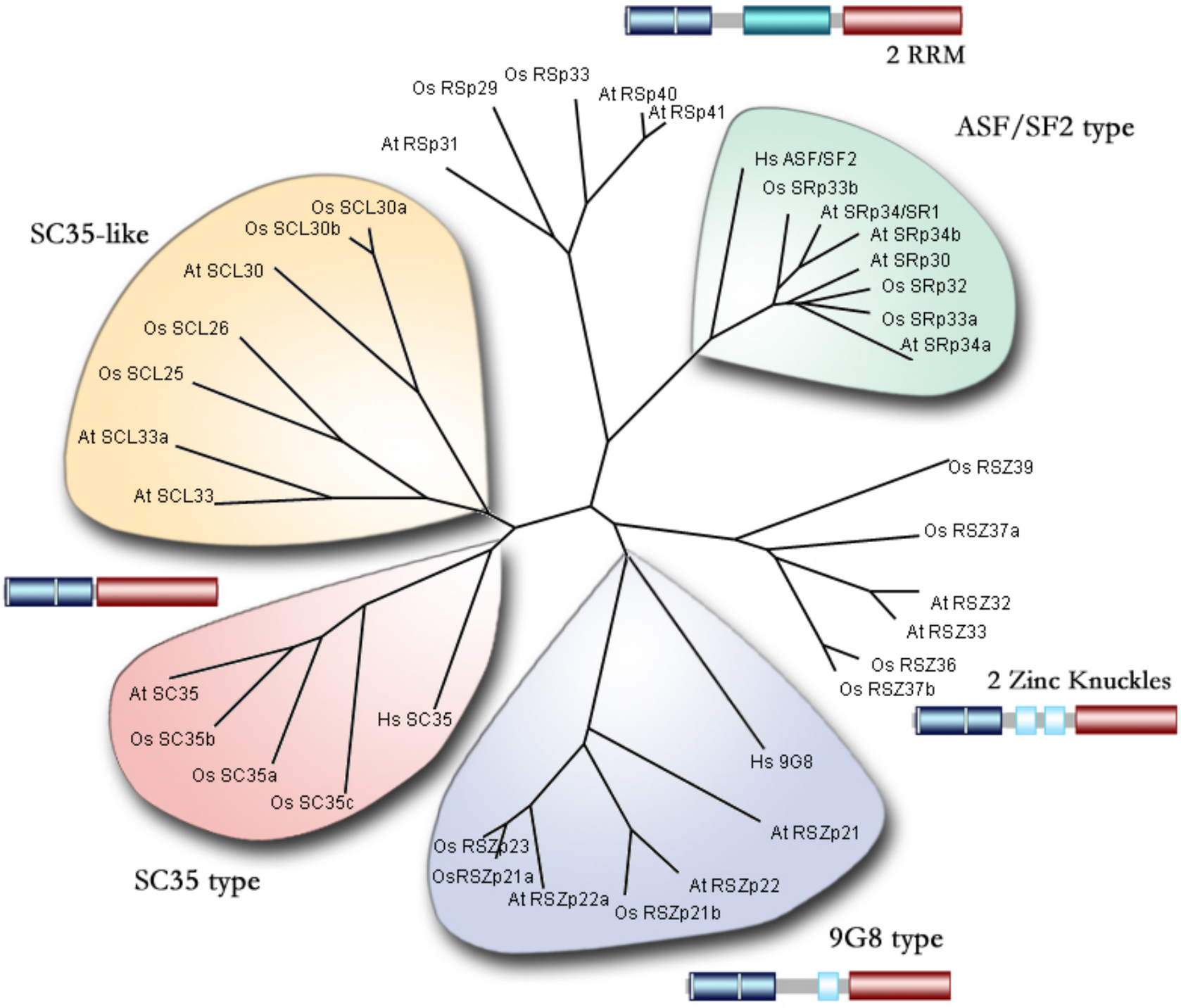
SR proteins are modular proteins displaying multiple architectures.

Nonetheless, they are often viewed as a natural protein family based on structural and functional criteria.

Isshiki et al. (2006) *Plant Cell* 18:146-158

Open Questions

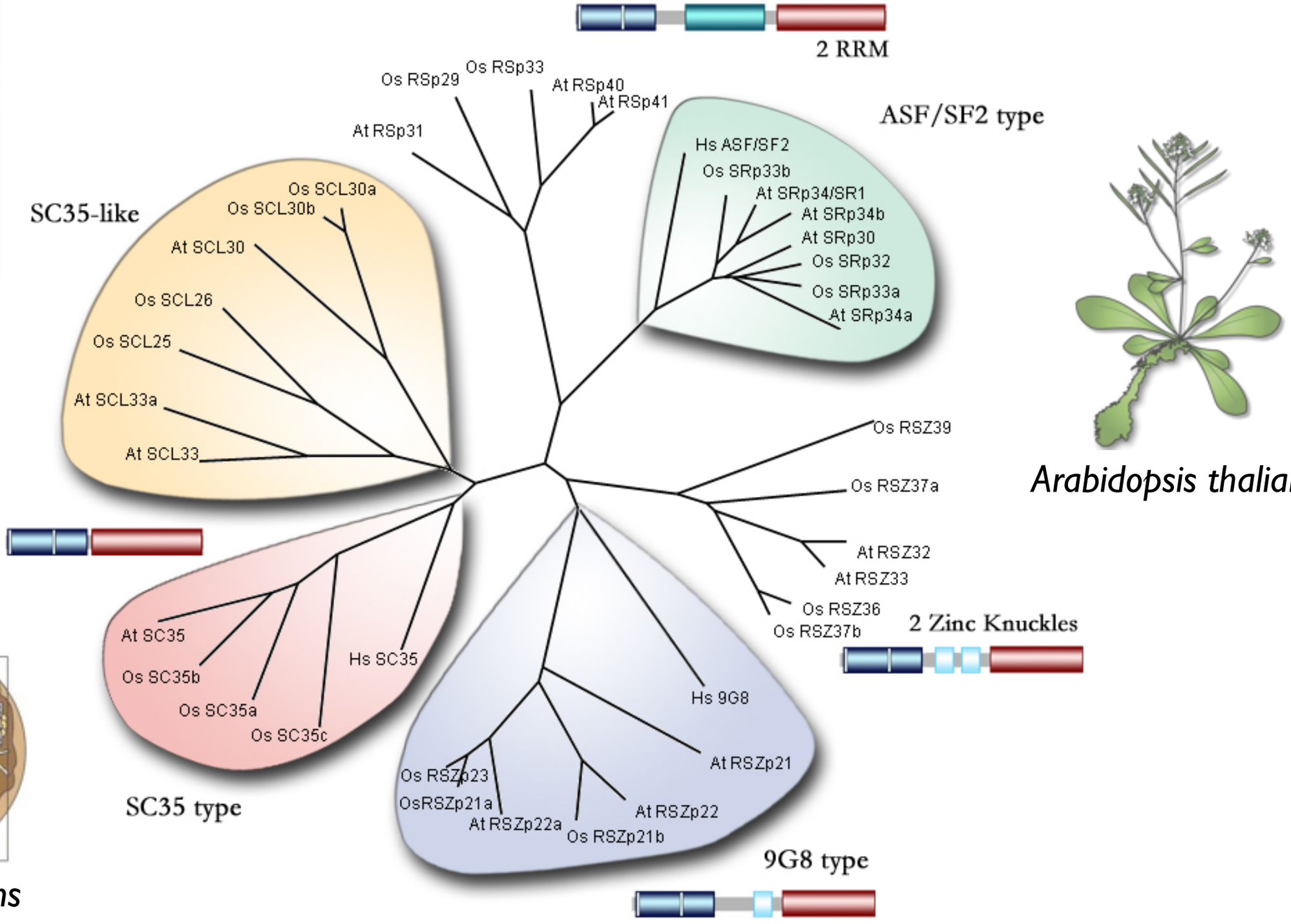
1. *Do all SR splicing factors indeed belong to a natural protein family?*
   (Do they all share a common origin?)
2. *How and when did they diversify to yield the currently recognized architectures?*
   (What are the relationships between animal and plant SR proteins?)



Some studies have tried to address these questions but often without adequate methodology...

modified from Isshiki et al. (2006) *Plant Cell* 18:146-158

Oryza sativa

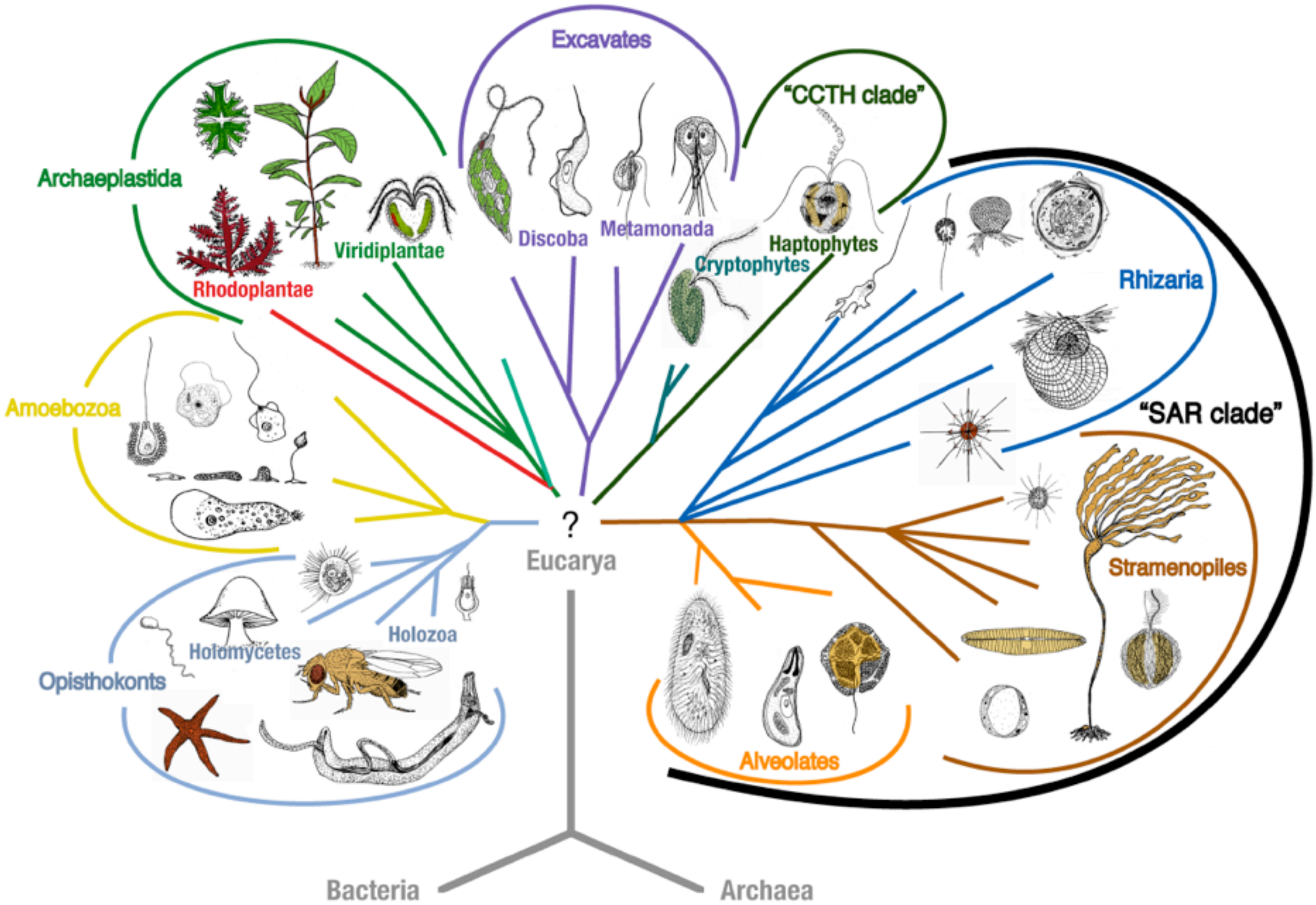


Arabidopsis thaliana

Homo sapiens

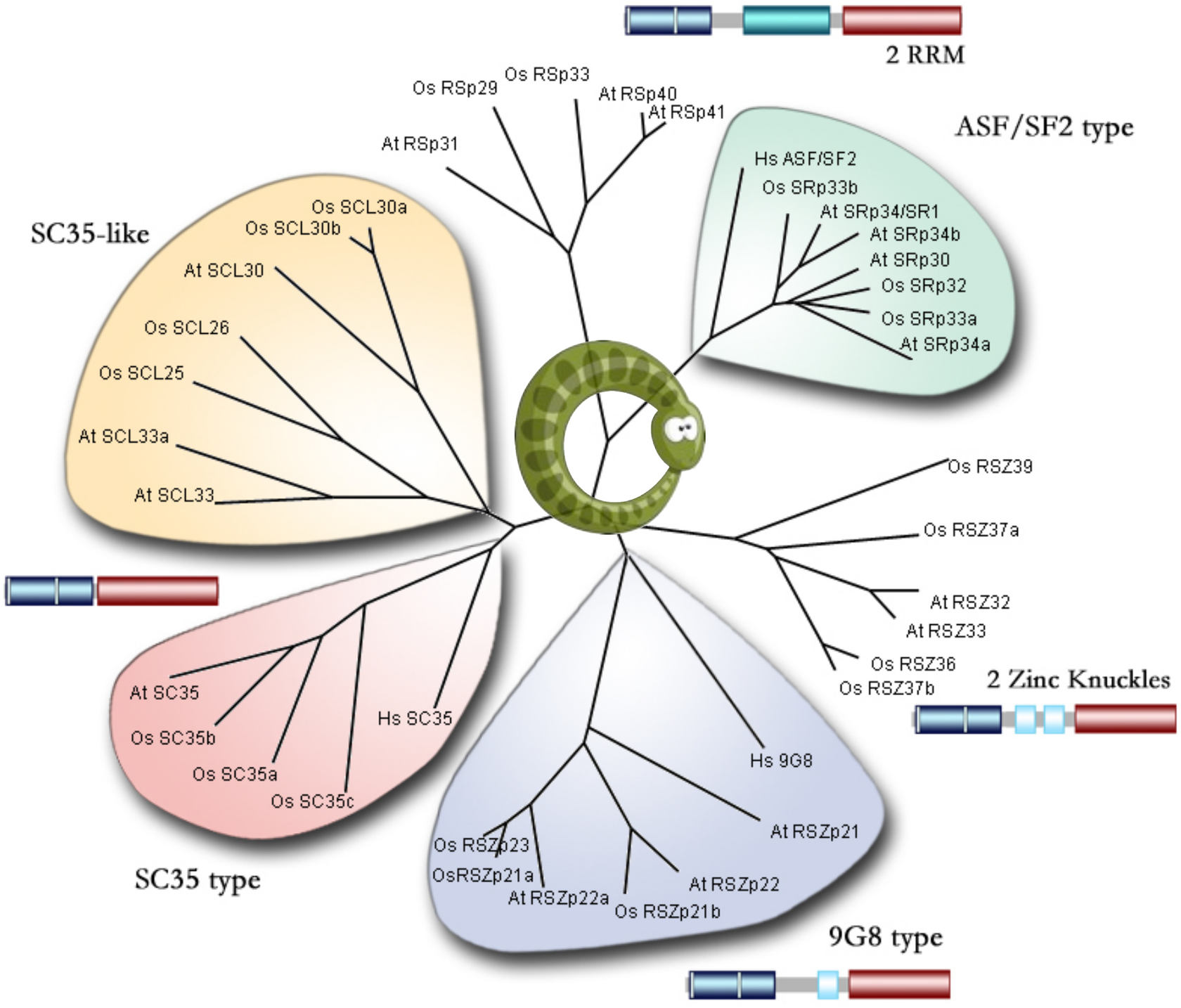
[1] Extremely sparse species sampling (here, only human, rice and Arabidopsis proteins)

modified from Isshiki et al. (2006) *Plant Cell* 18:146-158



Life is actually much more diverse!

modified from Walker et al. (2011) *Parasitology* 138:1638-1663

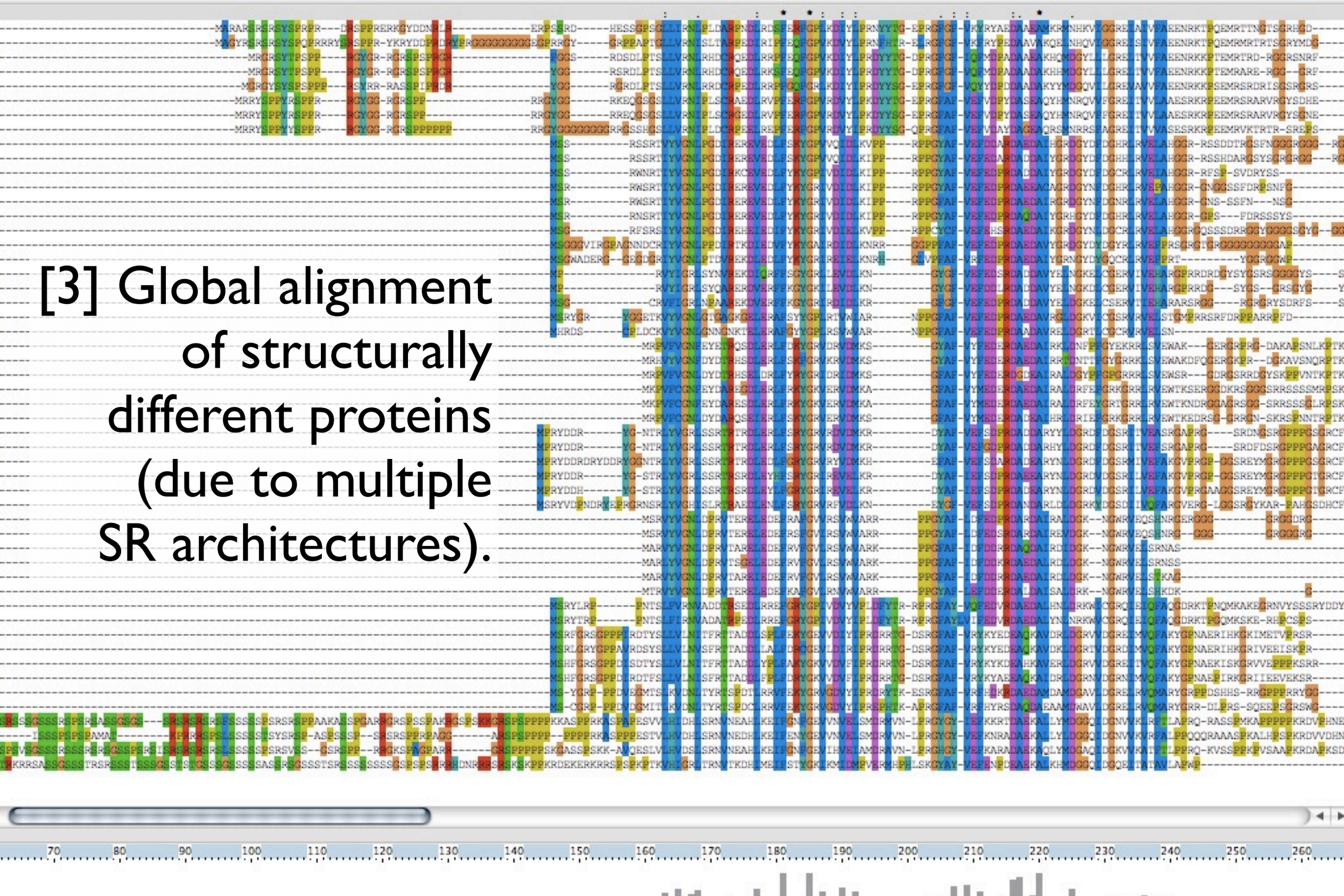


[2] « Expert » selection of sequences based on previous knowledge (no test of the common origin).

modified from Isshiki et al. (2006) *Plant Cell* 18:146-158

RRM1

[3] Global alignment of structurally different proteins (due to multiple SR architectures).



RRM2 / ZnK / RS

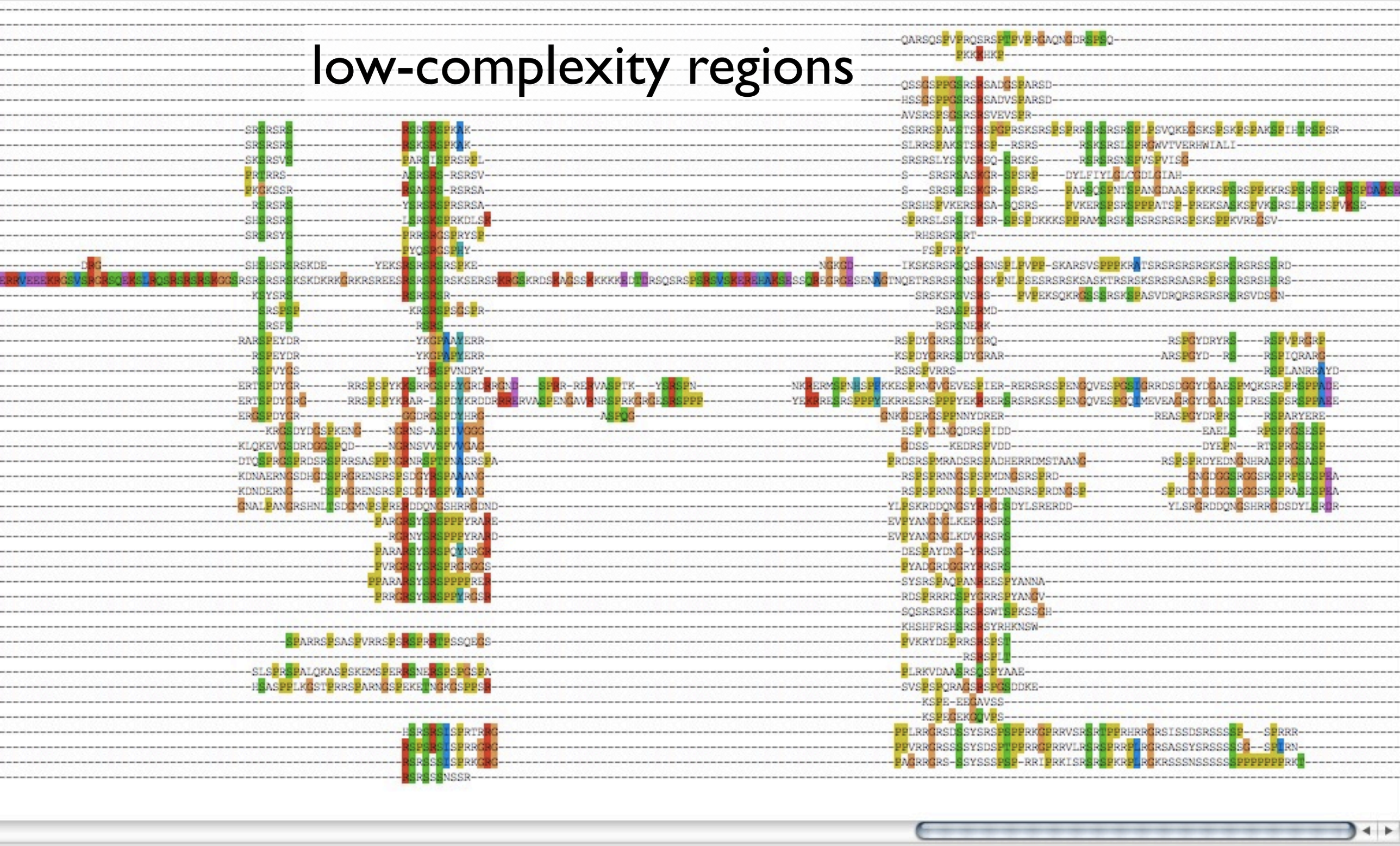
RS

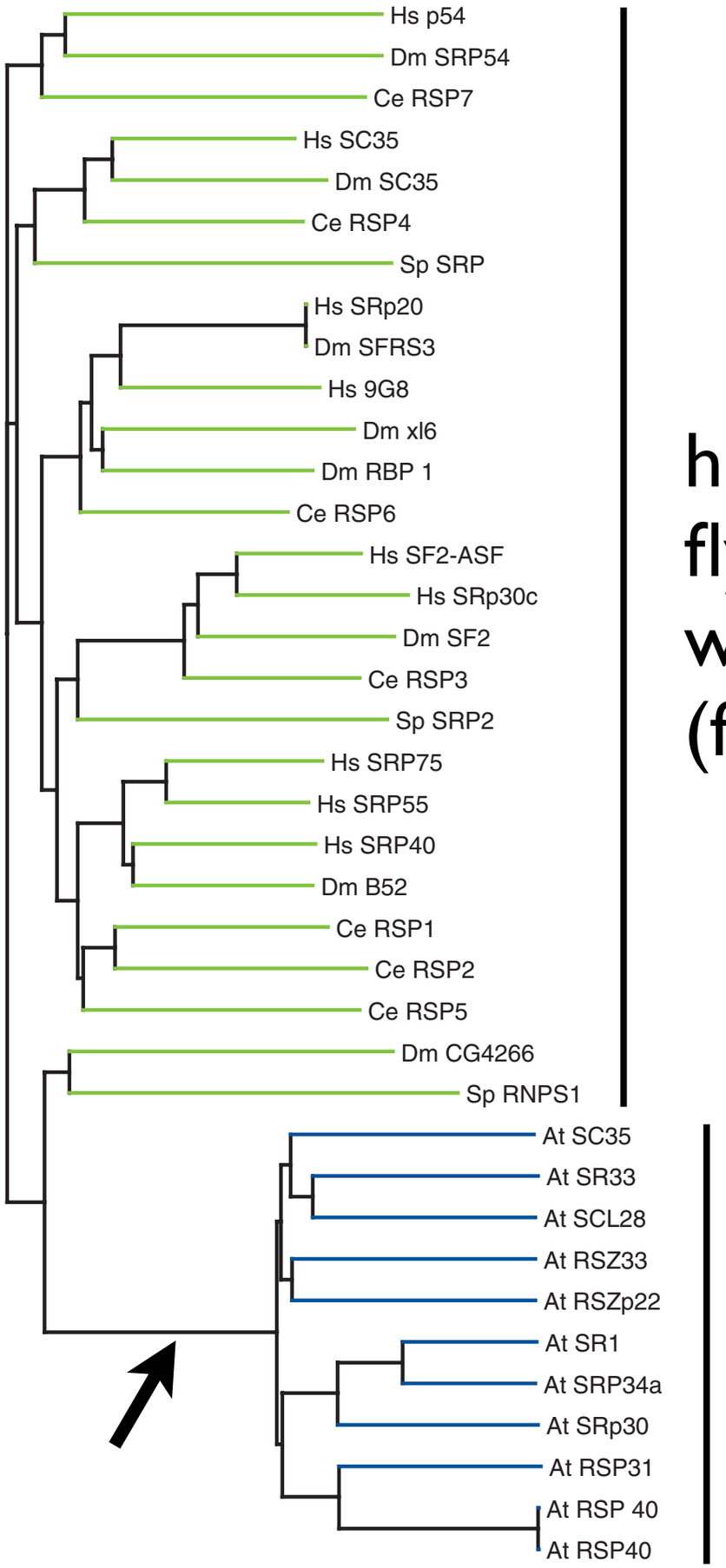
non-homologous regions

RS

RS

low-complexity regions





human
fly
worm
(fission) yeast

Arabidopsis

[4] Inadequate tree building methods leading to obvious phylogenetic artifacts.

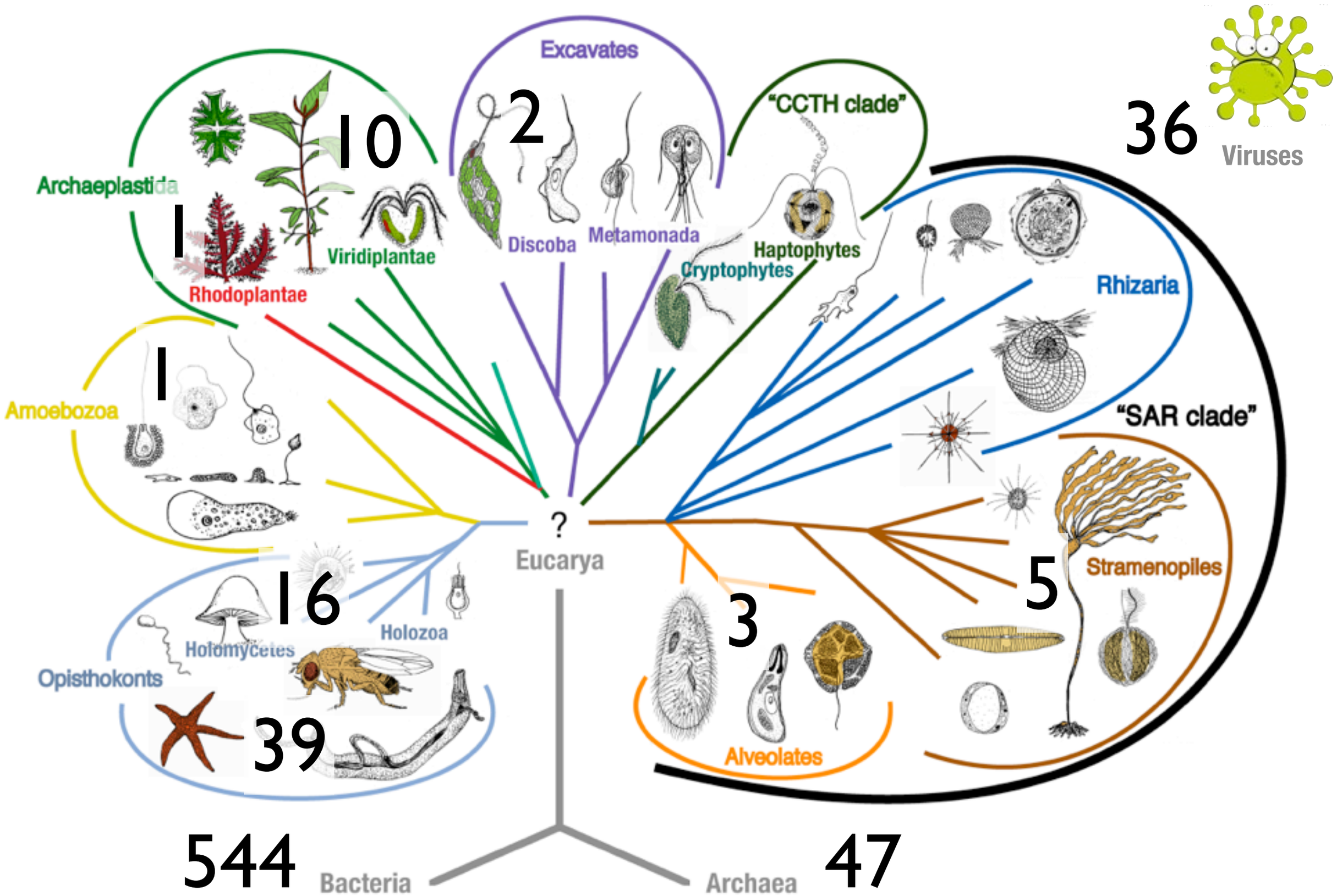
Here, plant SR proteins appear as a distinct group derived from a subset of animal/fungal proteins.

modified from Shepard and Hertel (2009) *Genome Biol* 10:242

Strategic Overview

1. extremely sparse species sampling
2. sequence selection (lack of outgroup)
3. global alignment of multiple architectures
4. inadequate tree building methods

- ~700 complete proteomes
- genome-wide pHMM predictions
- main focus on the RRM domains
- 2-step phylogenetic analysis with multiple approaches for better corroboration



[1] We sampled ~700 proteomes from the 3 domains.

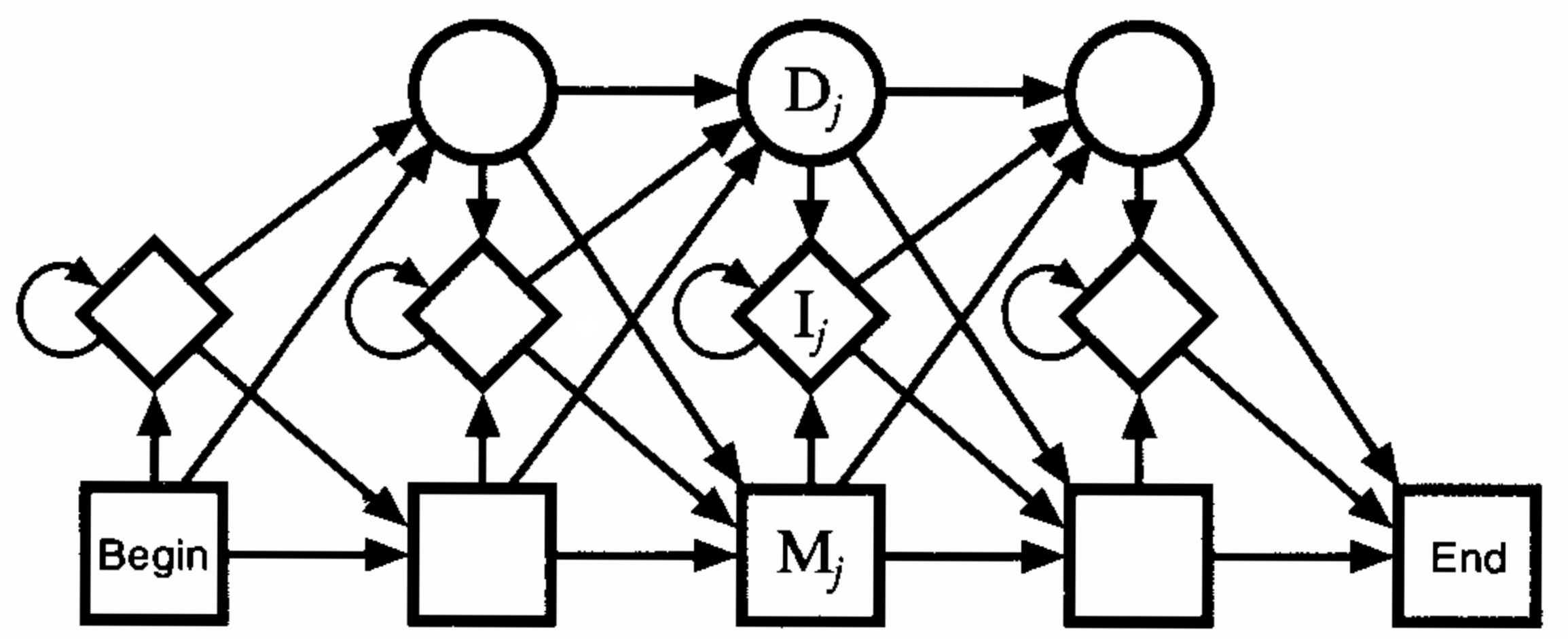
modified from Walker et al. (2011) *Parasitology* 138:1638-1663

Ensembl

NCBI

JGI

others



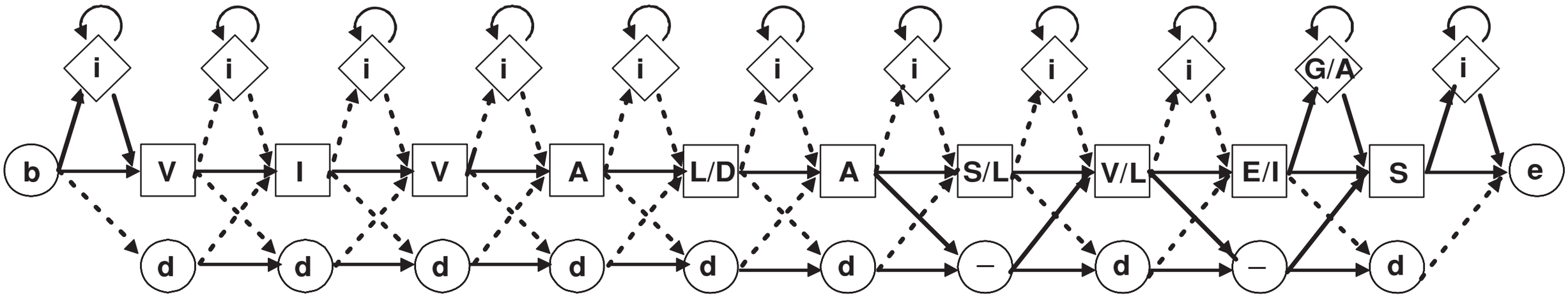
[2] We mined proteomes with a profile HMM (pHMM).

Durbin et al. (1998) *Biological Sequence Analysis. Probabilistic Models of Proteins and Nucleic Acids*

multiple alignment

```
V  I  V  A  L  A  S  V  E  G  A  S
V  I  V  A  D  A  -  V  I  -  -  S
V  I  V  A  D  A  L  L  -  -  A  S
```

pHMM



pHMMs summarize the features of a multiple protein alignment into a single model, against which novel sequences can be tested for similarity.

Cristianini and Hahn (2007) *Introduction to Computational Genomics*

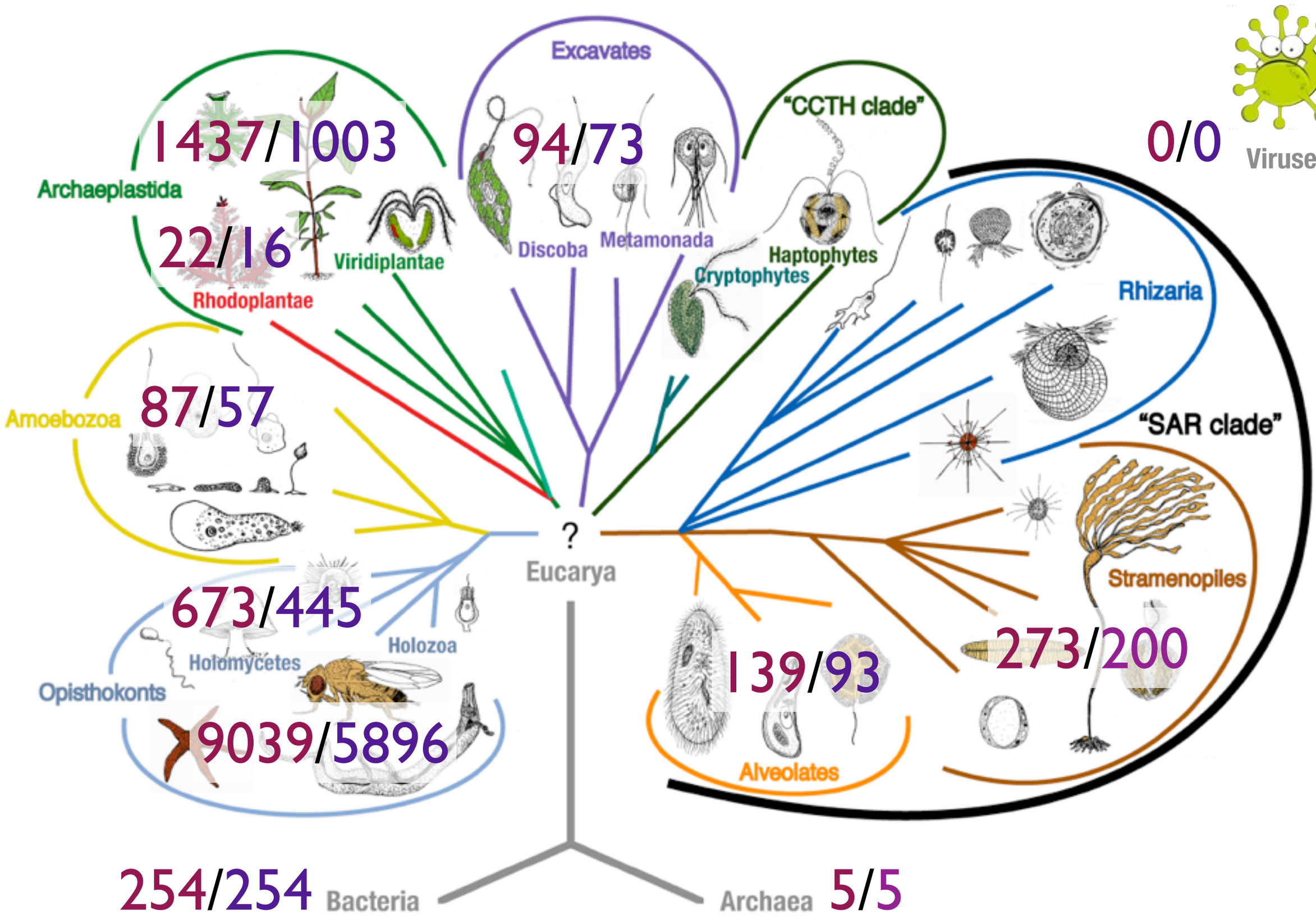
```
                                10        20        30        40        50        60        70
                       ....*....|....*....|....*....|....*....|....*....|....*....|....*....|....*....
gi 417810       39 IFIGNV--AHECTEDDLKQLFVEeFGDE-VSVEIPIKehtdGHIPASKHALVKFPTKIDFDNIKENYDT-KVVKDREIH 113
gi 17380398    177 VFVTNL--PNEYLHKDLVALFAK-FGRL-SALQRFTN------LNGNKSVLIAFDTSTGAEAVLQAKPKaLTLGDNVLS 245
gi 126947      197 LFVTNL--PRIVPYATLLELFSK-LGDV-KGIDTSSL-------STDGICIVAFFDIRQAIQAAKSLRSqRFFNDRLLY 264
gi 417689      200 IVIKFQ--GPALTEEEIYSLFRR-YGTI---IDIFP-----PTAANNNVAKVRYRSFRGAISAKNCVSG-IEIHNTVLH 266
gi 131528      186 IIVENL--FYPVTLDVLHQIFSK-FGTV---LKIIT-----FTKNNQFQALLQYADPVSAQHAKLSLDG-QNIYNACCT 252
gi 417742      229 LLIQNL--PSGTTEQLLSQILGN-EALV--EIRLVSV---------RNLAFVEYETVADATKIKNQLGS-TYKLQNNDV 292
gi 134095      153 LFLNNL--PEETNEMMLSMLFNQ-FPGF-KEVRLVPG--------RHDIAFVEFENDGQAGAARDALQGfKITPSHAMK 219
gi 48429205    453 LFVSGL--PMDAKPRELYLLFRA-YEGY-EGSLLKVts--kNGKTASPVGFVTFHTRAGAEAAKQDLQGvRFDPDMPQT 525
gi 465481      540 LYVGNL--PSDATEQELRQLFSG-QEGF-RRLSFRNKnttsNGHSHGPMCFVEFDDVSFATRALAELYG-RQLPRSTVS 613
gi 731605      348 IFVGQL--DKETTREELNRRFST-HGKI-QDINLIFK-------PTNIFAFIKYETEEAAAAALESENH-AIFLNKTMH 414
gi 1709755      43 VLVKNL--PKSYNQNKVYKYFKH-CGPI-IHVDVAD-----SLKKNFRFARIEFARYDGALAAIT-KTH-KVVGQNEII 110
gi 17380398    342 LVVENVgkHESYSSDALEKIFKK-FGDV-EEIDVVC--------SKAVLAFVTFKQSDAATKALAQLDG-KTVNKFEWK 409
gi 1709755     212 IMIRNLs-TELLDENLLRESFEG-FGSI-EKINIPAGq--kEHSFNNCCAFMVFENKDSAERALQ-MNR-SLLGNREIS 283
gi 1709851     373 LSVRNL--SPTVNNLLEEAFSQ-FGPI-ERAVVIVD---RGRSTGKGVEFAAKPAARKAFERCSE-GVFLLTTTP 442
gi 82592514    378 LRVSNL--TTFVSNLLYPHTIQ-EGPV-RASIVD---DRGTMGFGEFRSSASACLRMCNE-KCFFLTASL 447
gi 17380398    422 ILVTNL--TSDATEADLRKVFND-SGEV-ESIIMLG---------QRAVVKFKDDEGFCKSFL-ANE-SIVNNAPIF 483
gi 731637      315 ILVKNL--PSDTTQEEVLDYFST-IGPI-KSVFISEK-----QANTPHKAFVTYKNEEESKKAQKCLNK-TIFKNHTIW 383
gi 1172434     305 IFIKNL--PFETTRDDILNFFGE-VGPI-KSIYLSN-----AIKVKYLWAFVEYKNGSDSEKAIKRYNN-FYFRGKKLL 373
gi 206729905   216 VYCGGV--TSGITEQLMQSFSQ-FQPVMEIRVPD-----RGYSFVKFESHRSAAHAIVSVNG-TTIEGHVVK 280
gi 206729905     9 LYVGNL--SPDVTEALLQQFSQ-LGPVKNCKMIMD----RAGIPYCFVEFHRHAAAALAAMNG-RKIMGKEVK 77
gi 266991       16 LKVDNL--TYRTSPDTLRRVFEK-YGRV-GDVYIPRDr---YTKESRGFAFVRFHDKRDAEDAMDAMDG-AVLDGRELR 86
gi 23503093     10 LFIGGL--NTETNEKALEAVFGK-YGRI-VEVLLMKDr---ETNKSRGFAFVTFESPADAKDAARDMNG-KSLDGKAIK 80
gi 55584092    138 IFVGGL--TTEISDEEIKTYFGQ-FGNI-VEVEMPFDk---QKSQRKGFCFITFDSEQVVTDLLK-TPK-QKIAGKEVD 207
gi 172046785   107 IFVGGI--KEDTEEHHLRDYFEQ-YGKI-EVIEIMTDr---GSGKKRGFAFVTFDDHDSVDKIVI-QKY-HTVNGHNCE 176
gi 729000      161 IFVGGL--NPEATEEKIREYFGQ-FGEI-EAIELPIDp---KLNKRRGFVFITFKEEDPVKKVLE-KKF-HTVSGSKCE 230
gi 729000       77 MFVGGL--SWDTSKKDLKDYFTK-FGEV-VDCTIKMDp---NTGRSRGFGFILFKDSSSVEKVLD-QKE-HRLDGRVID 146
gi 55584092     58 LFVGGL--SWETTEKELRDHFGK-YGEI-ESINVKTDp---QTGRSRGFAFIVFTNTEAIDKVSA-ADE-HIINSKKVD 127
gi 400927       34 LFIGGL--APYTTEENLKLFYGQ-WGKV-VDVVVMRDa---ATKRSRGFGFITYTKSLMVDRAQE--NRpHIIDGKTVE 103
gi 136128       99 IGVFGL--NTNTSQHKVRELFNK-YGPI-ERIQMVIDa---QTQRSRGFCFIYFEKLSDARAAKDSCSG-IEVDGRRIR 169
gi 13638469    104 LFIARI--NYDTSESKLRREFEF-YGPI-KKIVLIHDq---ESGKPKGYAFIEYEHERDMHAAYKHADG-KKIDSKRVL 174
gi 544416        8 CFVGGL--AWATGDAELERTFSQ-FGEV-IDSKIINDr---ETGRSRGFGFVTFKDEKSMKDAIDEMNG-KELDGRTIT 78
gi 585573      292 VFVRNV--PYDATEESLAPHFSK-FGSV-KYALPVIDk---STGLAKGTAFVAFKDQYTYNECIKNAPA-AGSTSLLIG 362
gi 585573       28 LFVRSI--PQDVTDEQLADFFSN-FAPI-KHAVVVKD----TNKRSRGFGFVSFAVEDDTKEALAKARK-TKFNGHILR 97
gi 140346      124 IFVRNL--TFDCTPEDLKELFGT-VGEV-VEADIIT-----SKGHHRGMGTVEFTKNESVQDAISKFDG-ALFMDRKLM 192
...
```

79 seqs x 72 AA

E-value ≤ 1e-10

[3] We used a universal pHMM of the RRM domain.

http://www.ncbi.nlm.nih.gov/Structure/cdd/cddsrv.cgi?uid=pfam00076



HMMER predicted 12,023 RRMs from 8,042 proteins.

modified from Walker et al. (2011) *Parasitology* 138:1638-1663

◉ *Do SR proteins belong to a natural family?*

1. clustering based on sequence similarity
2. alignment of cluster representatives
3. tree reconstruction using multiple methods & models

◉ *How and when did SR proteins diversify?*

1. selection of SR-associated subtrees
2. alignment of all included domains
3. tree reconstruction using multiple methods, models & sequence samplings

[4] We carried out extensive phylogenetic analyses of the 12,023 RRM domains using a 2-step strategy, each one addressing a different question.

Do SR proteins belong to a natural family?

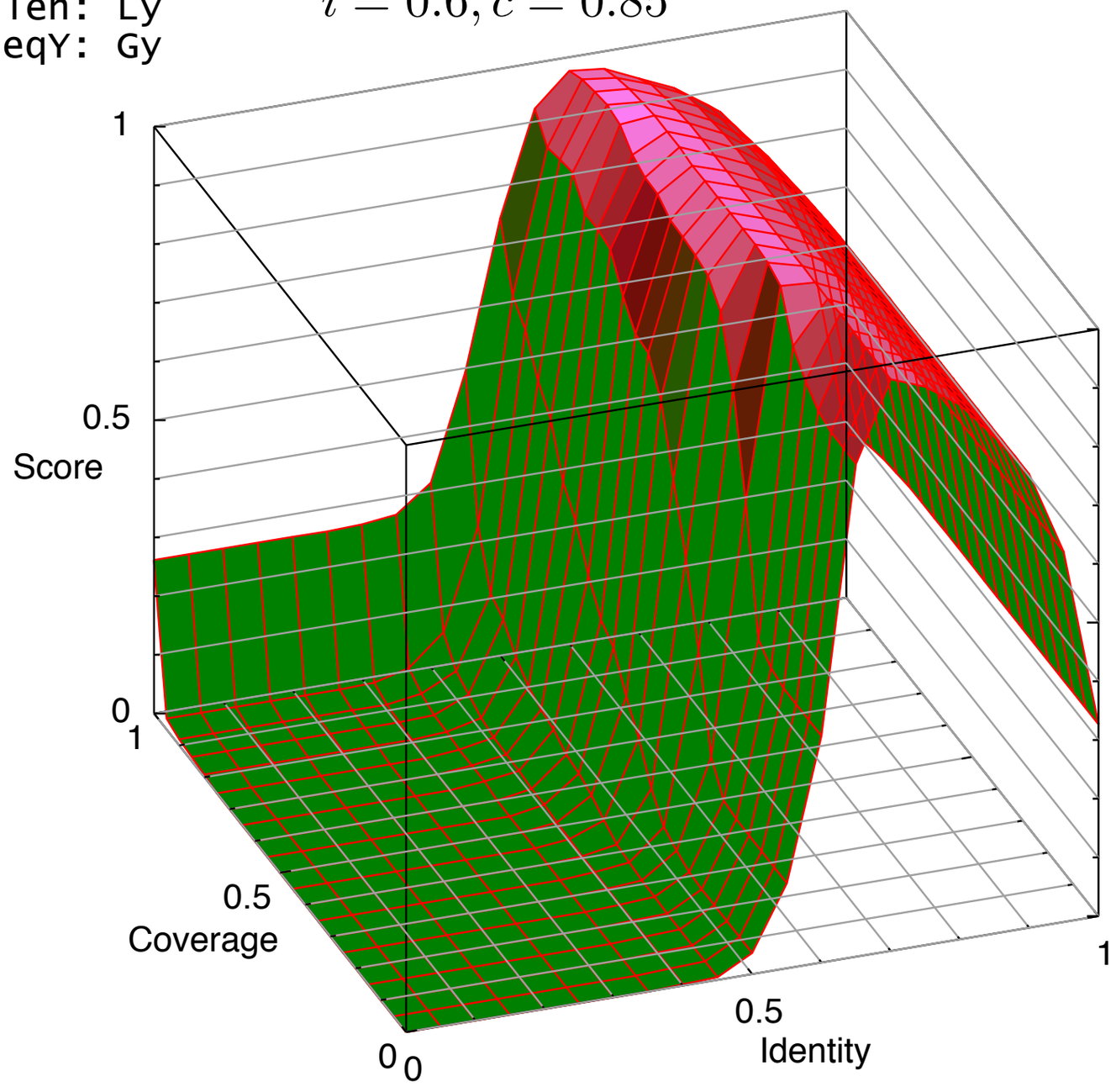
```
       x1.........Hx.........x2
        |                     |   gaps in seqX: Gx
seqX ---=======================-----  seqX len: Lx
         \\||||||||||||||||||//        matches: N
seqY ----=======================------  seqY len: Ly
        |                     |    gaps in seqY: Gy
       y1........Hy.......y2

coverage of seqX: Cx = Hx/Lx
coverage of seqY: Cy = Hy/Ly
coverage:         min(Cx,Cy)
alignment length  Al = Hx+Gx = Hy+Gy
identity:         N/Al*100%
```

$$H(C_k) = -\sum_{k=1}^{K} \frac{N_k}{N} \log \frac{N_k}{N}$$

$$Sc = \frac{H(C_k)}{\log K} - 0.8\frac{N_1}{N}$$

norm. entropy F(singl.)



Identical domains were merged and the resulting unique sequences were clustered based on similarity.

Califice et al. (2012) *Plant Physiol* 158:546-560

12,023

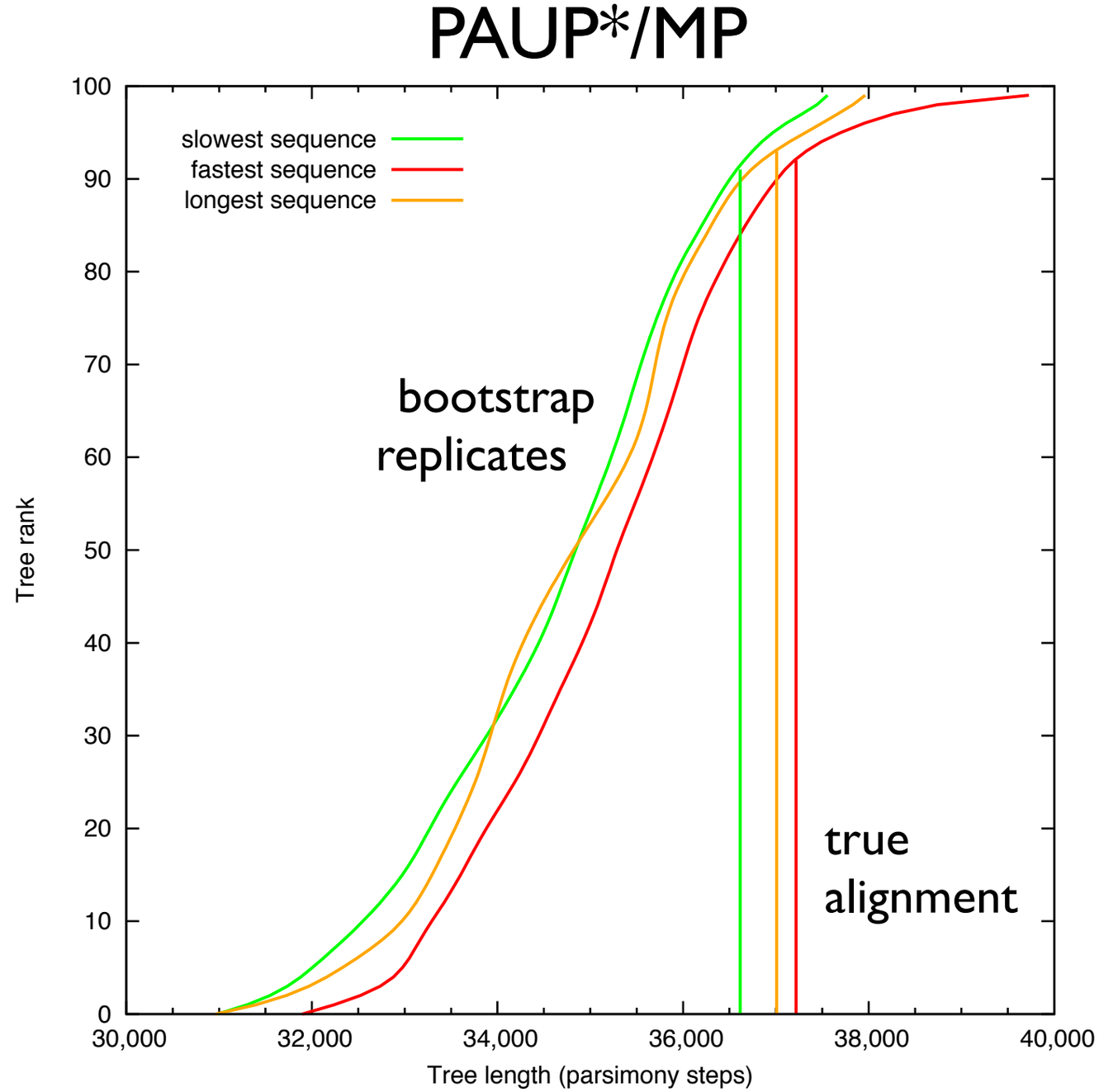
exact clustering

6633

similarity clustering

1266

$$S_x^* = \arg\max_x \sum_{y=1}^{J} \frac{\text{ident}(S_x, S_y)}{J}$$



Within each cluster, the slowest-evolving sequence was selected as the cluster representative.

Califice et al. (2012) *Plant Physiol* 158:546-560

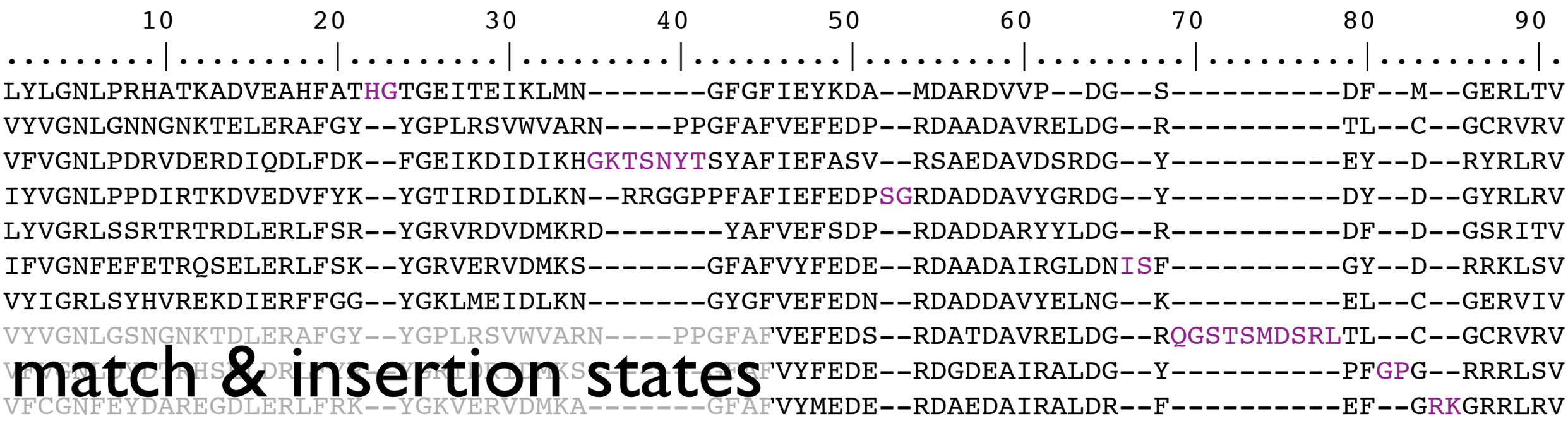
1. **Bos taurus@ENSBTAP00000012693_RRM_1**
   cellular organisms; Eukaryota; Opisthokonta; Metazoa; Eumetazoa; Bilateria; Coelomata; Deuterostomia; Chordata; Craniata; Vertebrata; Gnathostomata; Teleostomi; Euteleostomi; Sarcopterygii; Tetrapoda; Amniota; Mammalia; Theria; Eutheria; Laurasiatheria; Cetartiodactyla; Ruminantia; Pecora; Bovidae; Bovinae; Bos

2. **Canis familiaris@ENSCAFP00000017237_RRM_1**
   cellular organisms; Eukaryota; Opisthokonta; Metazoa; Eumetazoa; Bilateria; Coelomata; Deuterostomia; Chordata; Craniata; Vertebrata; Gnathostomata; Teleostomi; Euteleostomi; Sarcopterygii; Tetrapoda; Amniota; Mammalia; Theria; Eutheria; Laurasiatheria; Carnivora; Caniformia; Canidae; Canis; Canis lupus

3. **Equus caballus@XP_001503501_RRM_1**
   cellular organisms; Eukaryota; Opisthokonta; Metazoa; Eumetazoa; Bilateria; Coelomata; Deuterostomia; Chordata; Craniata; Vertebrata; Gnathostomata; Teleostomi; Euteleostomi; Sarcopterygii; Tetrapoda; Amniota; Mammalia; Theria; Eutheria; Laurasiatheria; Perissodactyla; Equidae; Equus; Equus subg. Equus

► **Laurasiatheria@ENSBTAP00000012693_RRM_1**

Cluster representatives were named after the most inclusive NCBI taxon including all cluster sequences.

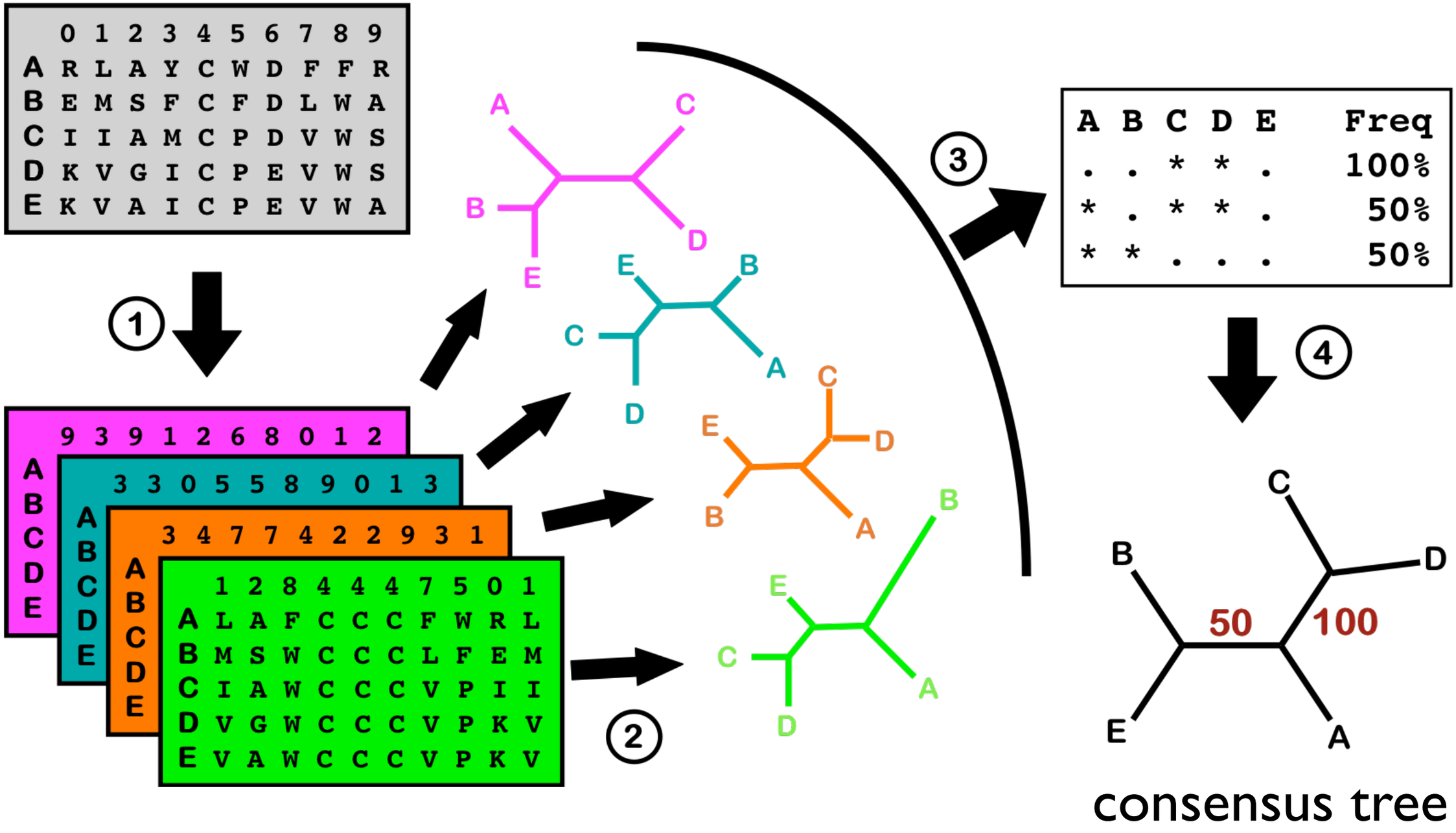


match & insertion states

only match states

1266 seqs
x 72 AA

Cluster representatives were aligned on the pHMM and insertions were discarded to limit alignment width.



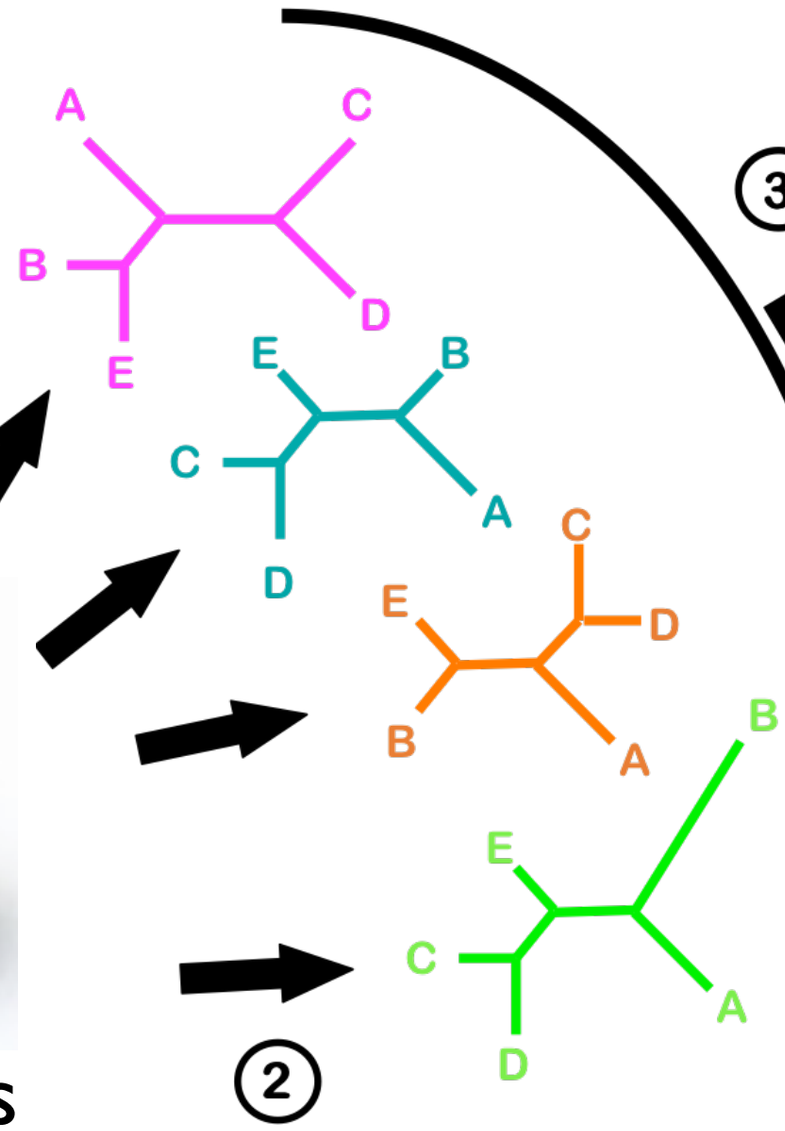
Statistical support assessment through data resampling (bootstraps) does not work with short sequences.

modified from Rodriguez-Ezpeleta (2007) *Ph.D Thesis*

	0	1	2	3	4	5	6	7	8	9
A	R	L	A	Y	C	W	D	F	F	R
B	E	M	S	F	C	F	D	L	W	A
C	I	I	A	M	C	P	D	V	W	S
D	K	V	G	I	C	P	E	V	W	S
E	K	V	A	I	C	P	E	V	W	A

1

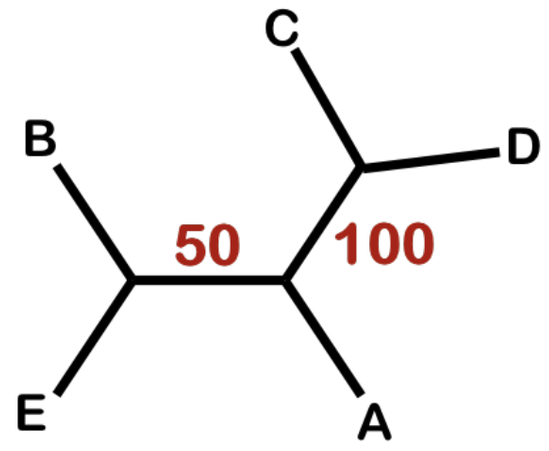
multiple methods & models



3

A	B	C	D	E	Freq
.	.	*	*	.	100%
*	.	*	*	.	50%
*	*	.	.	.	50%

4



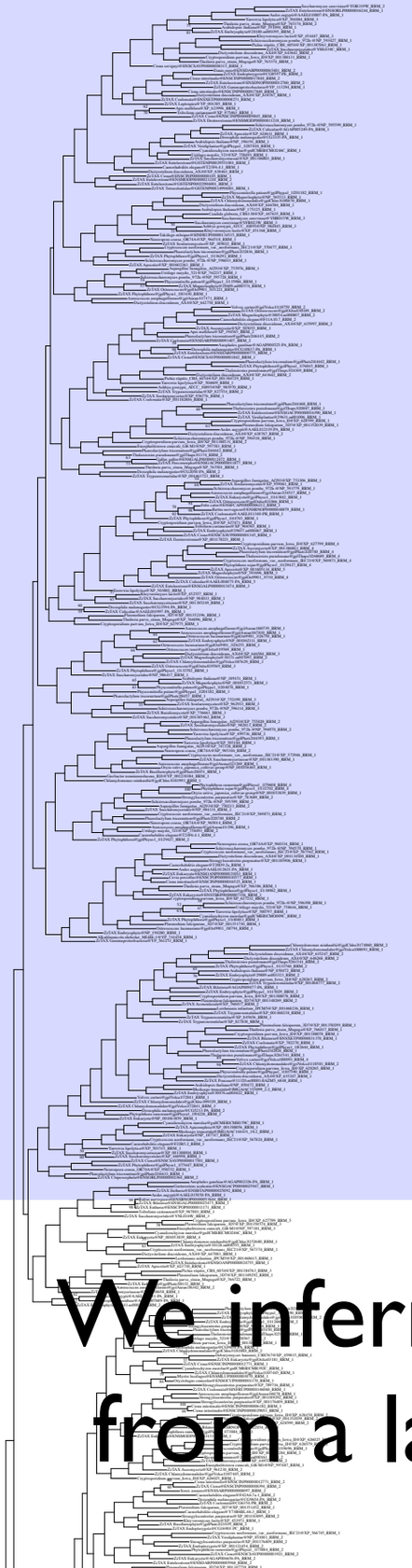
consensus tree

Instead, we did go for corroboration but did not use a strict consensus that would have been uninformative.

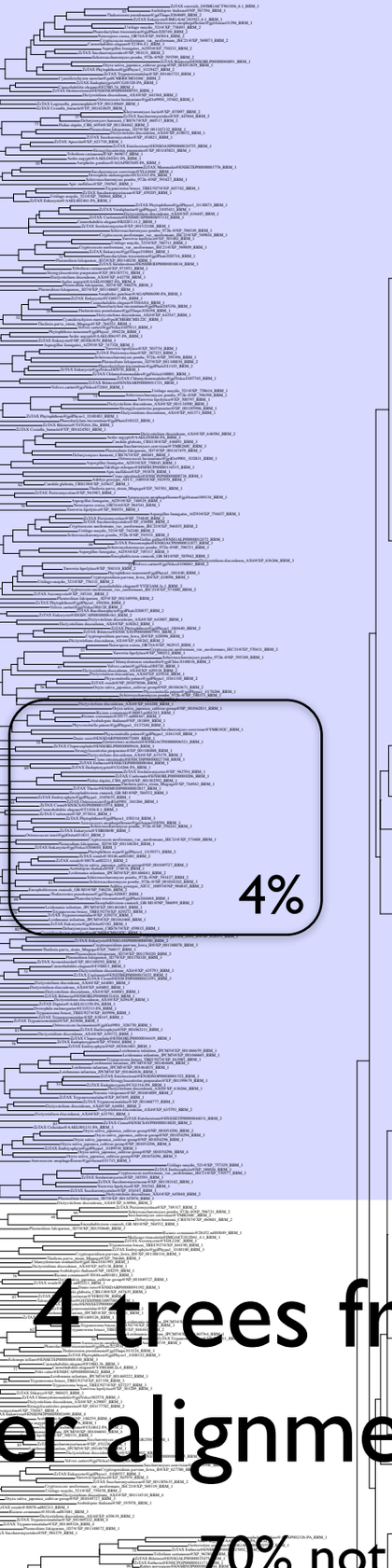
modified from Rodriguez-Ezpeleta (2007) *Ph.D Thesis*

1266x72

PAUP*/MP

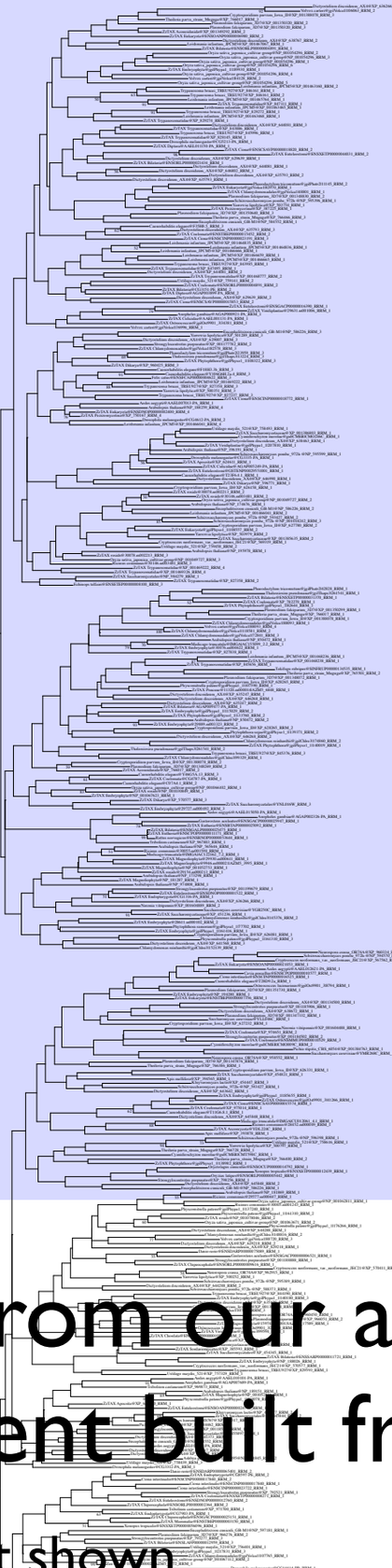


RAxML/WAG

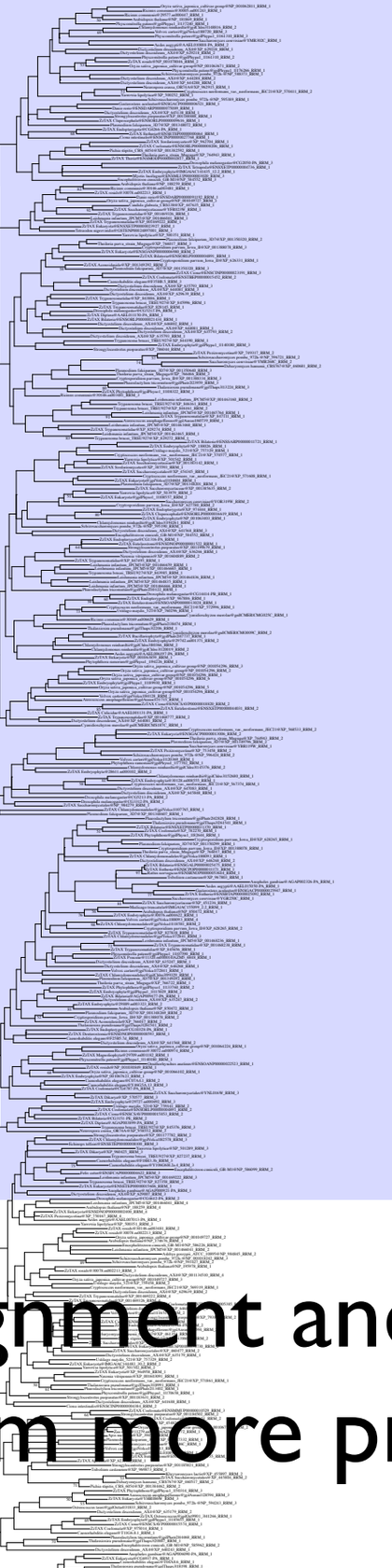


4%

TF/WAG

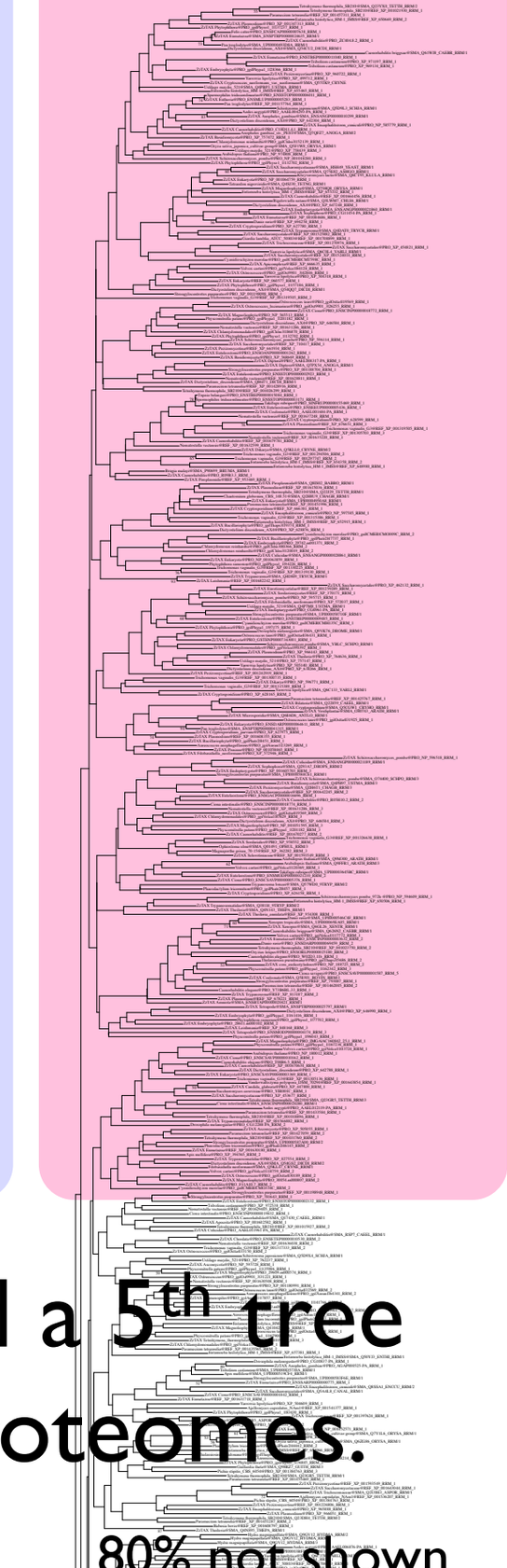


RAxML/LG



1831x72

RAxML/WAG



We inferred 4 trees from our alignment and a 5th tree from a larger alignment built from more proteomes.

70% not shown

80% not shown

Dictyostelium discoideum_AX4@XP_644288_RRM_1
Oryza sativa_japonica_cultivar-group@NP_001062811_RRM_1
Ricinus communis@30005.m001243_RRM_1
Ricinus communis@29577.m000447_RRM_1
Arabidopsis thaliana@NP_181869_RRM_1
Physcomitrella patens@jgi|Phypa1_1|137240_RRM_1
Saccharomyces cerevisiae@YMR302C_RRM_1
Physcomitrella patens@jgi|Phypa1_1|161310_RRM_3
Danio rerio@ENSDARP00000075089_RRM_1
55
Gasterosteus aculeatus@ENSGACP00000006521_RRM_1
ZzTAX Clupeocephala@ENSORLP00000009616_RRM_1
Strongylocentrotus purpuratus@XP_001188888_RRM_1
Dictyostelium discoideum_AX4@XP_635179_RRM_2
Ciona intestinalis@ENSCINP00000027768_RRM_1
ZzTAX Eutheria@ENSETEP00000000464_RRM_1
69
ZzTAX Endopterygota@CG4266-PA_RRM_1
ZzTAX Sordariomycetes@XP_962704_RRM_1
ZzTAX Coelomata@ENSORLP00000008206_RRM_1
Pichia stipitis_CBS_6054@XP_001382592_RRM_1
Theileria parva_strain_Muguga@XP_764943_RRM_1
ZzTAX Theria@ENSMODP00000002817_RRM_1
Encephalitozoon cuniculi_GB-M1@NP_584552_RRM_3
ZzTAX Embryophyta@jgi|Phypa1_1|185655_RRM_1
61
ZzTAX Ostreococcus@jgi|Ost9901_3|41266_RRM_1
ZzTAX Ciona@ENSCSAVP00000015574_RRM_1
Caenorhabditis elegans@T11G6.8.1_RRM_1
ZzTAX Coelomata@XP_975014_RRM_1
ZzTAX Phytophthora@jgi|Phyra1_1|50314_RRM_3
Aureococcus anophagefferens@jgi|Auran1|28594_RRM_3
Schizosaccharomyces pombe_972h-@NP_594243_RRM_3
ZzTAX Eukaryota@YHR086W_RRM_3
Ostreococcus tauri@jgi|Ostta4|31833_RRM_2
Cryptococcus neoformans_var._neoformans_JEC21@XP_571608_RRM_1
Plasmodium falciparum_3D7@XP_001348201_RRM_1
ZzTAX Eukaryota@jgi|Volca1|104604_RRM_1
Phytophthora sojae@jgi|Physo1_1|139371_RRM_2
ZzTAX rosids@30146.m003481_RRM_2
ZzTAX rosids@30078.m002213_RRM_2
Oryza sativa_japonica_cultivar-group@NP_001049727_RRM_2
Arabidopsis thaliana@NP_174676_RRM_1
Leishmania infantum_JPCM5@XP_001466041_RRM_2
Schizosaccharomyces pombe_972h-@NP_593427_RRM_2
83
Schizosaccharomyces pombe_972h-@NP_001018242_RRM_3
Ashbya gossypii_ATCC_10895@NP_984845_RRM_2
Encephalitozoon cuniculi_GB-M1@NP_586226_RRM_2
Thalassiosira pseudonana@jgi|Thaps3|20887_RRM_1
68
Phaeodactylum tricornutum@jgi|Phatr2|44468_RRM_1
Encephalitozoon cuniculi_GB-M1@NP_586099_RRM_2
Leishmania infantum_JPCM5@XP_001463465_RRM_1
78
Trypanosoma brucei_TREU927@XP_829272_RRM_1
ZzTAX Trypanosomatidae@XP_829274_RRM_1
Leishmania infantum_JPCM5@XP_001463468_RRM_1
ZzTAX Eukaryota@jgi|Ostta4|3181_RRM_1
82
Debaryomyces hansenii_CBS767@XP_459815_RRM_1

Thus, we had to find a way to compare these 5 trees by eye.

We annotated each leaf based on the content of the underlying RRM cluster.

Dictyostelium discoideum_AX4@XP_644288[RRM1][KOG3598]
Oryza sativa_japonica_cultivar-group@NP_001062811[RRM1][KOG1190]
Ricinus communis@30005_m001243[RRM1]
Ricinus communis@29577_m000447[RRM1][KOG0123]
Arabidopsis thaliana@NP_181869[4][RRM1:100%*][FCAFPA:100%*][KOG0117:100%*]
Physcomitrella patens@jgi_Phypa1_1_137240[2][RRM1:100%][KOG3598:50%]
Saccharomyces cerevisiae@YMR302C[RRM1]
Physcomitrella patens@jgi_Phypa1_1_161310[RRM3][KOG3598]
Danio rerio@ENSDARP00000075089[2][RRM1:100%][KOG0548:100%]
55
Gasterosteus aculeatus@ENSGACP00000006521[RRM1][KOG0548]
Clupeocephala@ENSORLP00000009616[3][RRM1:100%*][KOG0548:100%*]
Strongylocentrotus purpuratus@XP_001188888[2][RRM1:100%][KOG0548:100%]
Dictyostelium discoideum_AX4@XP_635179[RRM2][KOG0148]
Ciona intestinalis@ENSCINP00000027768[RRM1][KOG0548]
Eutheria@ENSETEP00000000464[10][RRM1:100%][put-sr:100%**][KOG0132:100%**]**
69
Endopterygota@CG4266-PA[5][RRM1:100%*][put-sr:100%*][KOG0132:100%*]
Sordariomycetes@XP_962704[2][RRM1:100%][put-sr:50%][KOG4206:100%]
Coelomata@ENSORLP00000008206[18][RRM1:100%][snRNP:83%**][KOG4206:100%**]**
Pichia stipitis_CBS_6054@XP_001382592[RRM1][KOG4206]
Theileria parva_strain_Muguga@XP_764943[RRM1][KOG4206]
Theria@ENSMODP00000002817[2][RRM1:100%]
Encephalitozoon cuniculi_GB-M1@NP_584552[RRM3][KOG0123]
Embryophyta@jgi_Phypa1_1_185655[7][RRM1:100%*][KOG0153:100%*]
61
Ostreococcus@jgi_Ost9901_3_41266[2][RRM1:100%][KOG0153:100%]
Ciona@ENSCSAVP00000015574[2][RRM1:100%][KOG0153:100%]
Caenorhabditis elegans@T11G6_8_1[2][RRM1:100%][KOG0153:100%]
Coelomata@XP_975014[28][RRM1:100%*][RBM:86%***][KOG0153:100%***]**
Phytophthora@jgi_Phyra1_1_50314[2][RRM3:100%][KOG0148:100%]
Aureococcus anophagefferens@jgi_Auran1_28594[RRM3][KOG0148]
Schizosaccharomyces pombe_972h-@NP_594243[RRM3][KOG0148]
Eukaryota@YHR086W[51][RRM3:76%*][.RRM2:20%**][oligoU:10%*][KOG0148:90%***]**
Ostreococcus tauri@jgi_Ostta4_31833[RRM2][KOG0148]
Cryptococcus neoformans_var._neoformans_JEC21@XP_571608[RRM1][put-sr][KOG0151]
Plasmodium falciparum_3D7@XP_001348201[RRM1][KOG0151]
Eukaryota@jgi_Volca1_104604[49][RRM1:100%*][put-sr:78%***][KOG0151:100%***]**
Phytophthora sojae@jgi_Physo1_1_139371[RRM2][KOG4205]
rosids@30146_m003481[2][RRM2:100%][PABP:50%][KOG0123:100%]
rosids@30078_m002213[3][RRM2:100%*][PABP:33%][KOG0123:100%*]
Oryza sativa_japonica_cultivar-group@NP_001049727[RRM2][KOG0123]
Arabidopsis thaliana@NP_174676[RRM1][PABP][KOG0123]
Leishmania infantum_JPCM5@XP_001466041[RRM2][KOG0123]
Schizosaccharomyces pombe_972h-@NP_593427[RRM2][KOG0123]
83
Schizosaccharomyces pombe_972h-@NP_001018242[RRM3][KOG0123]
Ashbya gossypii_ATCC_10895@NP_984845[RRM2][KOG0123]
Encephalitozoon cuniculi_GB-M1@NP_586226[RRM2][put-sr][KOG0123]
Thalassiosira pseudonana@jgi_Thaps3_20887[RRM1]
68
Phaeodactylum tricornutum@jgi_Phatr2_44468[RRM1]
Encephalitozoon cuniculi_GB-M1@NP_586099[RRM2][KOG0131]
Leishmania infantum_JPCM5@XP_001463465[RRM1][KOG0122]
78
Trypanosoma brucei_TREU927@XP_829272[RRM1][KOG0145]
Trypanosomatidae@XP_829274[2][RRM1:100%][KOG0145:50%]
Leishmania infantum_JPCM5@XP_001463468[RRM1]
Eukaryota@jgi_Ostta4_3181[66][RRM1:100%**][put-sr:47%***][cyclophilin:32%***][KOG0415:97%****]**
82
Debaryomyces hansenii_CBS767@XP_459815[RRM1][KOG0415]

Then each leaf was color-coded according to its annotation.

Dictyostelium discoideum_AX4@XP_644288[RRM1][KOG3598]
Oryza sativa_japonica_cultivar-group@NP_001062811[RRM1][KOG1190]
Ricinus communis@30005_m001243[RRM1]
Ricinus communis@29577_m000447[RRM1][KOG0123]
Arabidopsis thaliana@NP_181869[4][RRM1:100%*][FCAFPA:100%*][KOG0117:100%*]
Physcomitrella patens@jgi_Phypa1_1_137240[2][RRM1:100%][KOG3598:50%]
Saccharomyces cerevisiae@YMR302C[RRM1]
Physcomitrella patens@jgi_Phypa1_1_161310[RRM3][KOG3598]
Danio rerio@ENSDARP00000075089[2][RRM1:100%][KOG0548:100%]
55
Gasterosteus aculeatus@ENSGACP00000006521[RRM1][KOG0548]
Clupeocephala@ENSORLP00000009616[3][RRM1:100%*][KOG0548:100%*]
Strongylocentrotus purpuratus@XP_001188888[2][RRM1:100%][KOG0548:100%]
Dictyostelium discoideum_AX4@XP_635179[RRM2][KOG0148]
Ciona intestinalis@ENSCINP00000027768[RRM1][KOG0548]
Eutheria@ENSETEP00000000464[10][RRM1:100%**][put-sr:100%**][KOG0132:100%**]
69
Endopterygota@CG4266-PA[5][RRM1:100%*][put-sr:100%*][KOG0132:100%*]
Sordariomycetes@XP_962704[2][RRM1:100%][put-sr:50%][KOG4206:100%]
Coelomata@ENSORLP00000008206[18][RRM1:100%**][snRNP:83%**][KOG4206:100%**]
Pichia stipitis_CBS_6054@XP_001382592[RRM1][KOG4206]
Theileria parva_strain_Muguga@XP_764943[RRM1][KOG4206]
40
Theria@ENSMODP00000002817[2][RRM1:100%]
Encephalitozoon cuniculi_GB-M1@NP_584552[RRM3][KOG0123]
Embryophyta@jgi_Phypa1_1_185655[7][RRM1:100%*][KOG0153:100%*]
61
Ostreococcus@jgi_Ost9901_3_41266[2][RRM1:100%][KOG0153:100%]
Ciona@ENSCSAVP00000015574[2][RRM1:100%][KOG0153:100%]
Caenorhabditis elegans@T11G6.8_[2][RRM1:100%][KOG0153:100%]
Coelomata@XP_975014[28][RRM1:100%***][RRM:86%***][KOG0153:100%***]
14 - RBM
Phytophthora@jgi_Phyra1_1_50314[2][RRM3:100%][KOG0148:100%]
Aureococcus anophagefferens@jgi_Auran1_28594[RRM3][KOG0148]
Schizosaccharomyces pombe_972h-@NP_594243[RRM3][KOG0148]
Eukaryota@YHR086W[51][RRM3:76%***][.RRM2:20%**][oligoU:10%*][KOG0148:90%***]
78 - OligoU
Ostreococcus tauri@jgi_Ostta4_31833[RRM2][KOG0148]
Cryptococcus neoformans_var._neoformans_JEC21@XP_571608[RRM1][put-sr][KOG0151]
Plasmodium falciparum_3D7@XP_001348201[RRM1][KOG0151]
Eukaryota@jgi_Volca1_104604[49][RRM1:100%***][put-sr:78%***][KOG0151:100%***]
13
Phytophthora sojae@jgi_Physo1_1_139371[RRM2][KOG4205]
rosids@30146_m003481[2][RRM2:100%][PABP:50%][KOG0123:100%]
rosids@30078_m002213[3][RRM2:100%*][PABP:33%][KOG0123:100%*]
Oryza sativa_japonica_cultivar-group@NP_001049727[RRM2][KOG0123]
Arabidopsis thaliana@NP_174676[RRM1][PABP][KOG0123]
Leishmania infantum_JPCM5@XP_001466041[RRM2][KOG0123]
Schizosaccharomyces pombe_972h-@NP_593427[RRM2][KOG0123]
83
Schizosaccharomyces pombe_972h-@NP_001018242[RRM3][KOG0123]
Ashbya gossypii_ATCC_10895@NP_984845[RRM2][KOG0123]
Encephalitozoon cuniculi_GB-M1@NP_586226[RRM2][put-sr][KOG0123]
Thalassiosira pseudonana@jgi_Thaps3_20887[RRM1]
68
Phaeodactylum tricornutum@jgi_Phatr2_44468[RRM1]
Encephalitozoon cuniculi_GB-M1@NP_586099[RRM2][KOG0131]
Leishmania infantum_JPCM5@XP_001463465[RRM1][KOG0122]
78
Trypanosoma brucei_TREU927@XP_829272[RRM1][KOG0145]
Trypanosomatidae@XP_829274[2][RRM1:100%][KOG0145:50%]
Leishmania infantum_JPCM5@XP_001463468[RRM1]
Eukaryota@jgi_Ostta4_3181[66][RRM1:100%****][put-sr:47%***][cyclophilin:32%***][KOG0415:97%****]
82
Debaryomyces hansenii_CBS767@XP_459815[RRM1][KOG0415]

Finally, colored trees were analyzed by my co-authors!

8042

RRM extraction

12,023

exact clustering

6633

similarity clustering

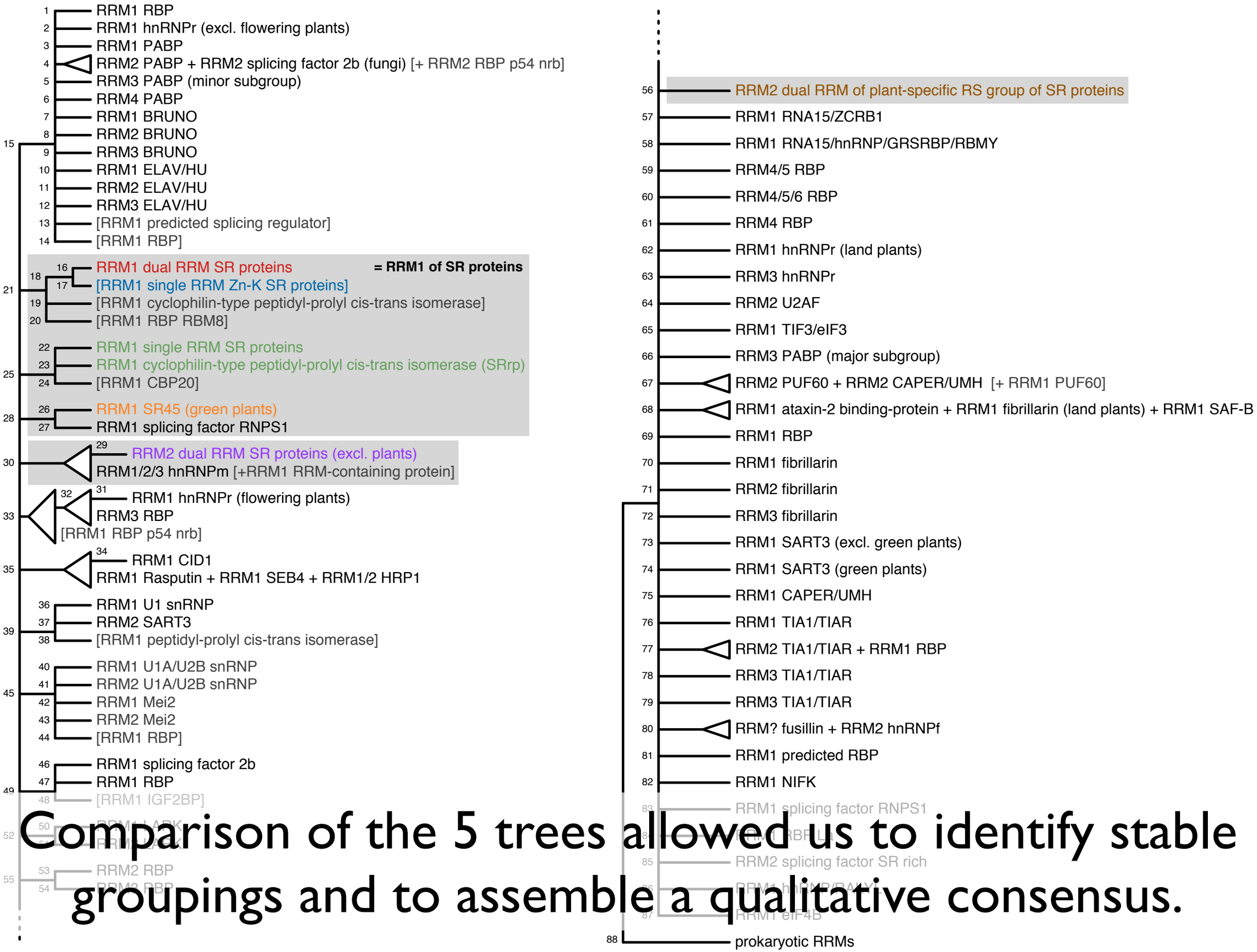
1266

1. protein annotation
   - KOG affiliation
   - similarity to reference RRM-containing proteins
   - prokaryotic source
   - presence of RS repeats

2. domain annotation
   - RRM rank in protein
3. cluster annotation
   - cluster size (in domains)

Annotation required keeping live links from 1266 leaves to 12,023 RRM domains and 8042 full-length proteins.

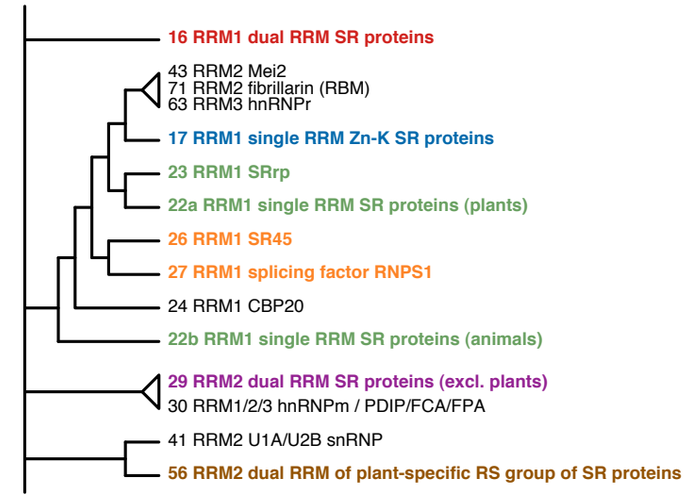
Furthermore, an efficient strategy had to be devised to summarize protein annotation at the cluster level.



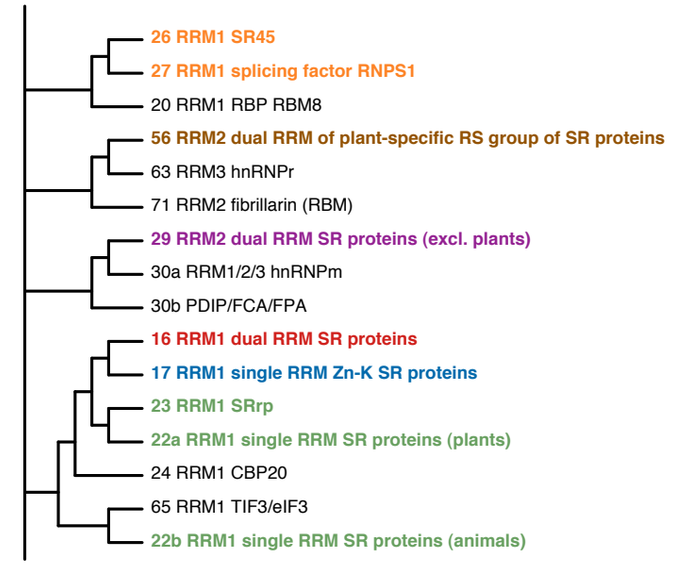
Comparison of the 5 trees allowed us to identify stable groupings and to assemble a qualitative consensus.

Califice et al. (2012) *Plant Physiol* 158:546-560

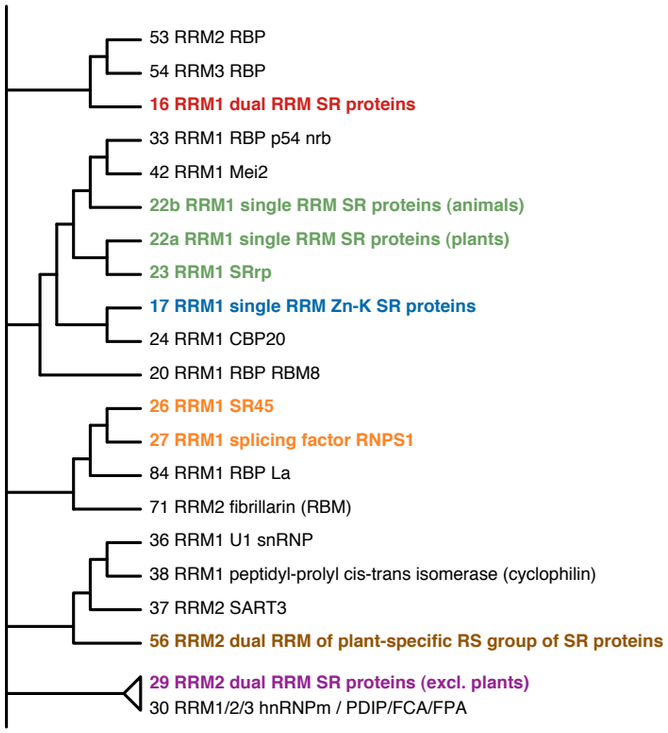
A PAUP* (MP)



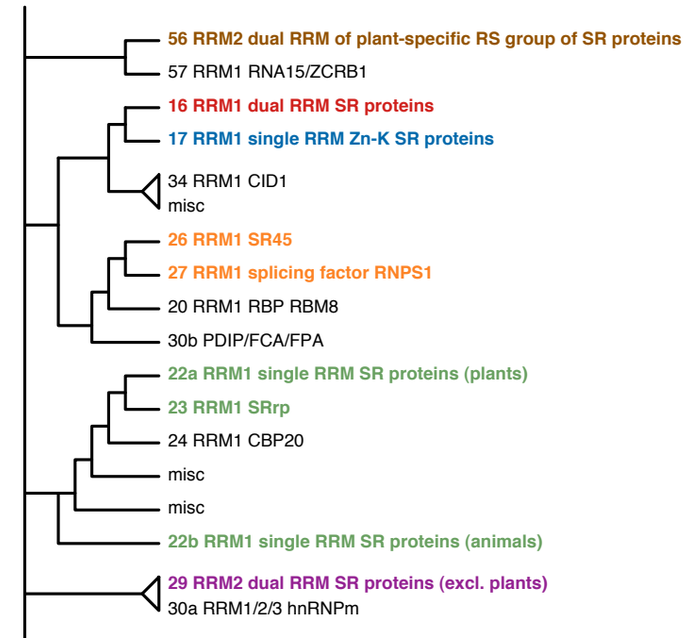
B RAxML (WAG+Γ_4)



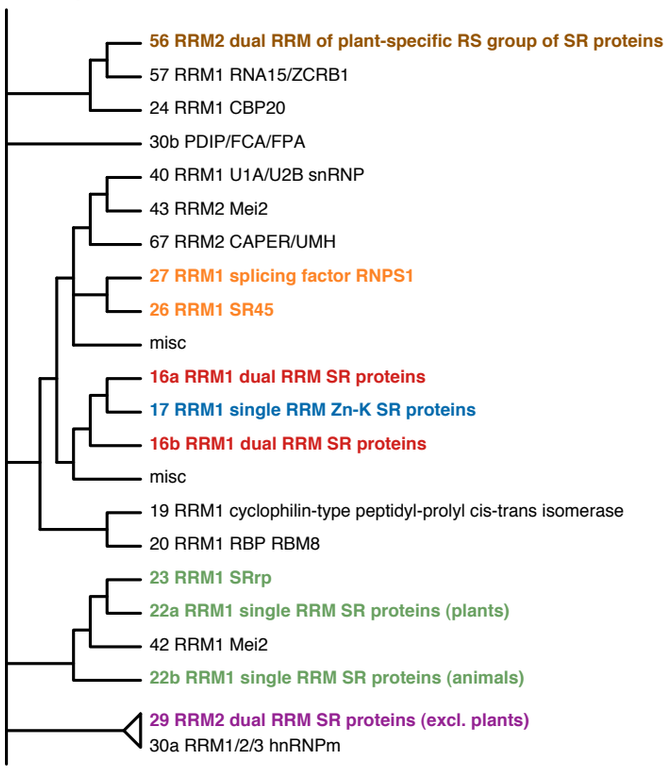
C TREEFINDER (WAG+Γ_4)



D RAxML (LG+F+Γ_4)



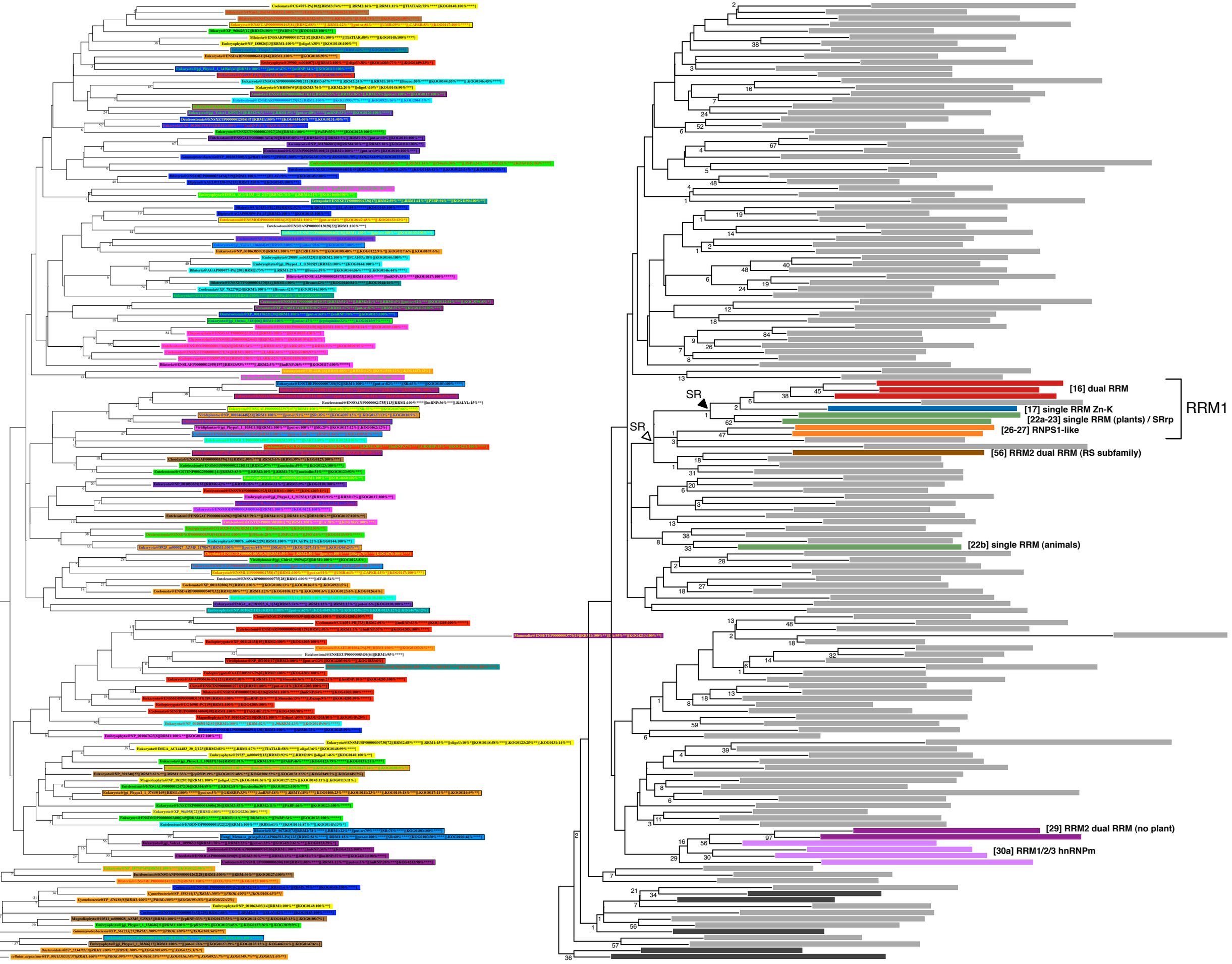
E large RAxML (WAG+Γ_4)

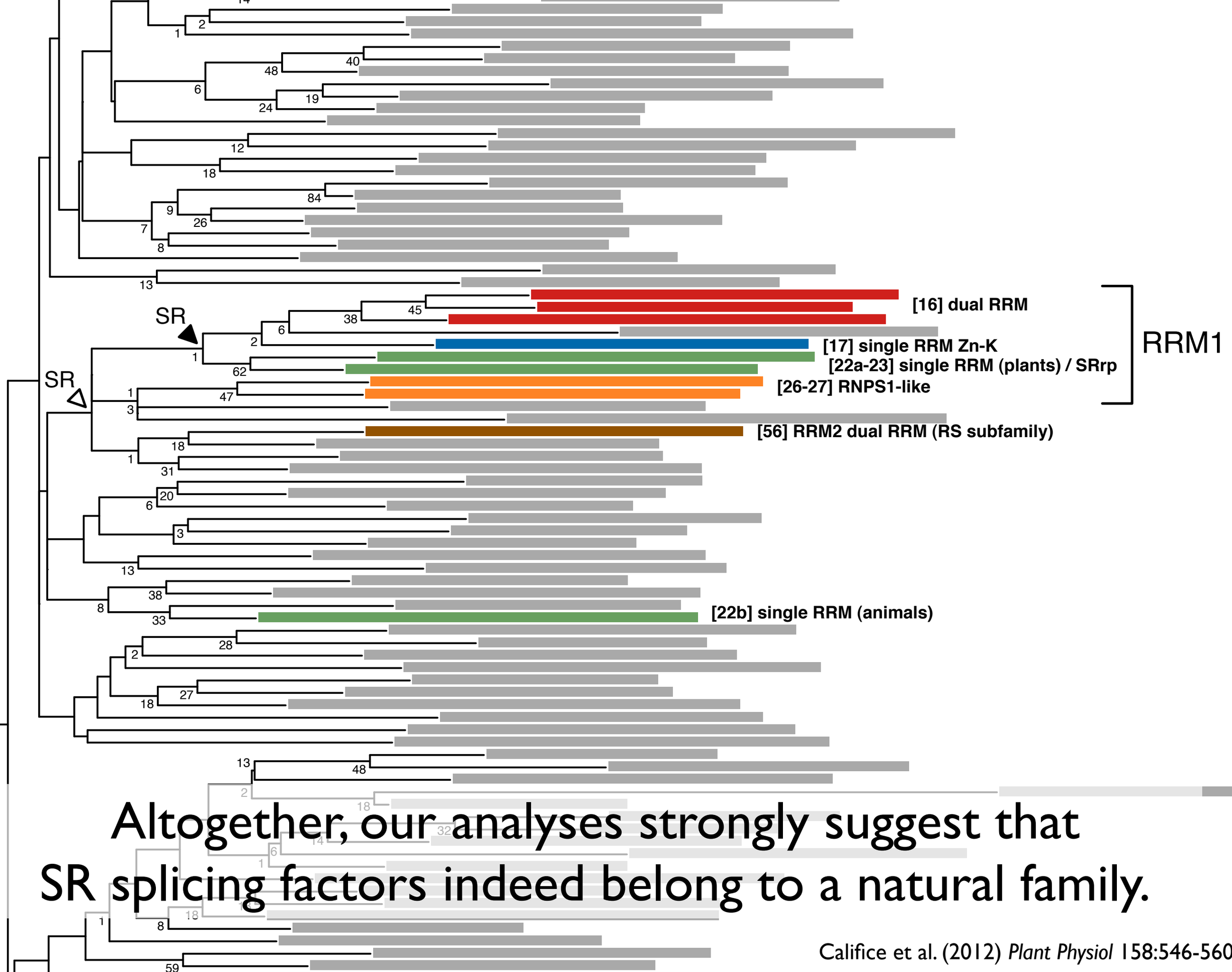


RRM1 of SR proteins fell into 4 groups of mutual affinities but were never recovered as a single group.

To further test a possible common origin, we built a tree based on the 152 largest clusters.

Califice et al. (2012) *Plant Physiol* 158:546-560

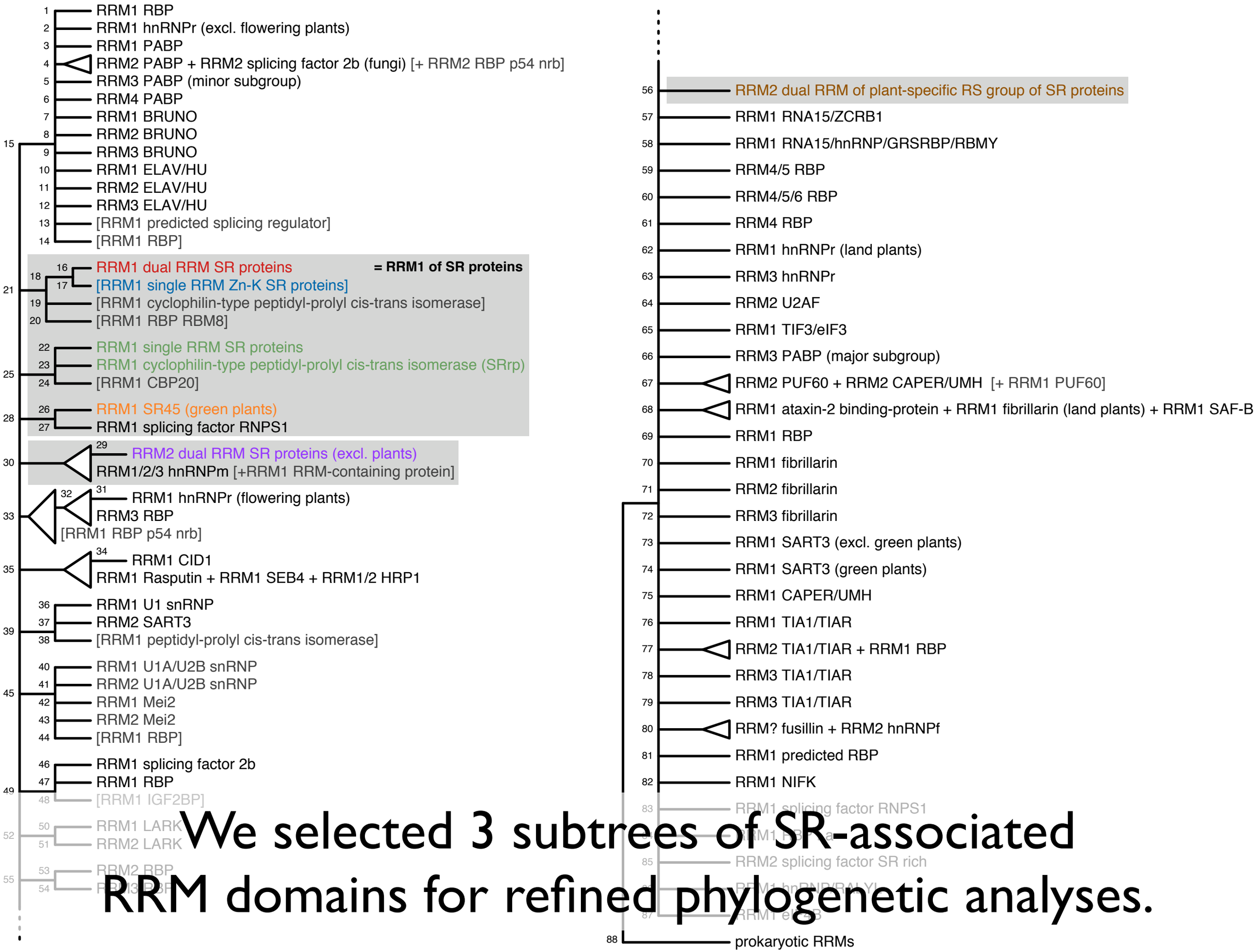

[16] dual RRM
[17] single RRM Zn-K
[22a-23] single RRM (plants) / SRrp
[26-27] RNPS1-like
RRM1
SR
SR
[56] RRM2 dual RRM (RS subfamily)
[22b] single RRM (animals)
[29] RRM2 dual RRM (no plant)
[30a] RRM1/2/3 hnRNPm



Altogether, our analyses strongly suggest that SR splicing factors indeed belong to a natural family.

Califice et al. (2012) *Plant Physiol* 158:546-560

How and when did SR proteins diversify?



We selected 3 subtrees of SR-associated RRM domains for refined phylogenetic analyses.

Califice et al. (2012) *Plant Physiol* 158:546-560

```
        10        20        30        40        50        60        70        80        90
.........|.........|.........|.........|.........|.........|.........|.........|.........|.
LYLGNLPRHATKADVEAHFATHGTGEITEIKLMN-------GFGFIEYKDA--MDARDVVP--DG--S----------DF--M--GERLTV
VYVGNLGNNGNKTELERAFGY--YGPLRSVWVARN----PPGFAFVEFEDP--RDAADAVRELDG--R----------TL--C--GCRVRV
VFVGNLPDRVDERDIQDLFDK--FGEIKDIDIKHGKTSNYTSYAFIEFASV--RSAEDAVDSRDG--Y----------EY--D--RYRLRV
IYVGNLPPDIRTKDVEDVFYK--YGTIRDIDLKN--RRGGPPFAFIEFEDPSGRDADDAVYGRDG--Y----------DY--D--GYRLRV
LYVGRLSSRTRTRDLERLFSR--YGRVRDVDMKRD-------YAFVEFSDP--RDADDARYYLDG--R----------DF--D--GSRITV
IFVGNFEFETRQSELERLFSK--YGRVERVDMKS-------GFAFVYFEDE--RDAADAIRGLDNISF----------GY--D--RRKLSV
VYIGRLSYHVREKDIERFFGG--YGKLMEIDLKN-------GYGFVEFEDN--RDADDAVYELNG--K----------EL--C--GERVIV
VYVGNLGSNGNKTDLERAFGY--YGPLRSVWVARN----PPGFAFVEFEDS--RDATDAVRELDG--RQGSTSMDSRLTL--C--GCRVRV
VFVGNLPYDTRHSDLRRLFSK--YGRIDDIYMKSG-------FGFVYFEDE--RDGDEAIRALDG--Y----------PFGPG--RRRLSV
VFCGNFEYDAREGDLERLFRK--YGKVERVDMKA-------GFAFVYMEDE--RDAEDAIRALDR--F----------EF--GRKGRRLRV
```

match & insertion states

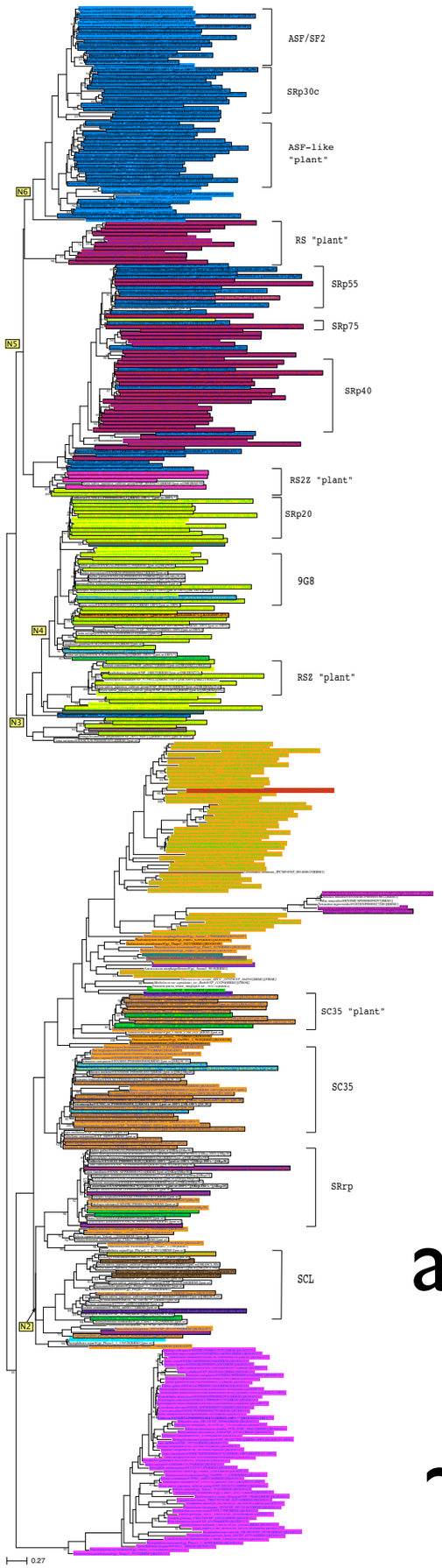
```
        10        20        30        40        50        60        70
.........|.........|.........|.........|.........|.........|.........|..
LFVGGINPDTDNEKVKEHFSQFGEIEEFERPVDRTTEKNRGFCFITFKKDGCIKLACA-QRNQELDGSKVDV
LTVKNLPRGTNPQDLHKLFKKFGTVTLCRITND--------QAVVHMRFPSMATNAVRNLSGEIYRGNVLSV
LCITQLPLAFTYDQFLSLITPFGTPERCFLVHSDVTGHSMGYGCVEFTSKESSIKAKNQLNGFKIQNNVLQV
IFVGDLDASVTDDKLEDFFLKYRSVKGAKIMYE-EGGVSRGYGFVRFSDESEQKRALVEMQGKGLGAKSIRV
IYVGSLSYDVTADELQSFFGQYGEIEEAKLIMDRETGRSKGFAFITYGTQDAAQEAVSKANGIDLQGRKIRV
IFVGRLAPSVDENQFRKLFERFGTIVRCRLVRDVVTELSRGYGFVEFSERKSALRAIDEMHGRNLEGKELLV
LNVSNVNFSTSKEEIEEHFRKAGRVKGVRIPKK----RASGFAFVEMLDAEGFQKAFL-LHGSFLDGRQIRV
LFLKNLSFQTNEESIRETFRNMGPIHSVQVVRRKGKNESRGYGFIQFKLRKSADSALKNLQSVHIDGRKVEL
LFFANLPYEITNESLRQIFGPFGAIRTVELPMDPKTNKHRGFGFVTFESPDSVDKVVEVRFHTINGSSGKSRGFGYVEFENEQDVQKALS-FDRRPLDGRPVFI
VFVGNIARGTTDNDLHEFFSRVGPIEYVRQIGD--------KYVAYVCFKKGVSIMKALK-LNQESLNGRLIRV
```

RRM1:
434 seqs
x 93 AA

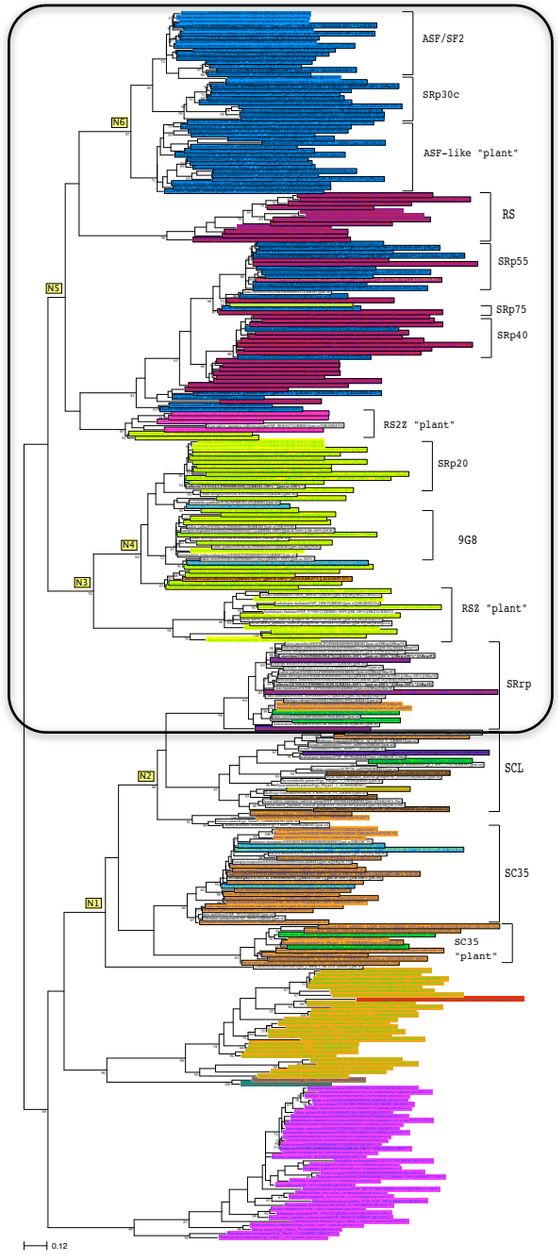
only match states

For this step, we aligned unclustered RRM domains and kept sequence insertions to maximize signal.

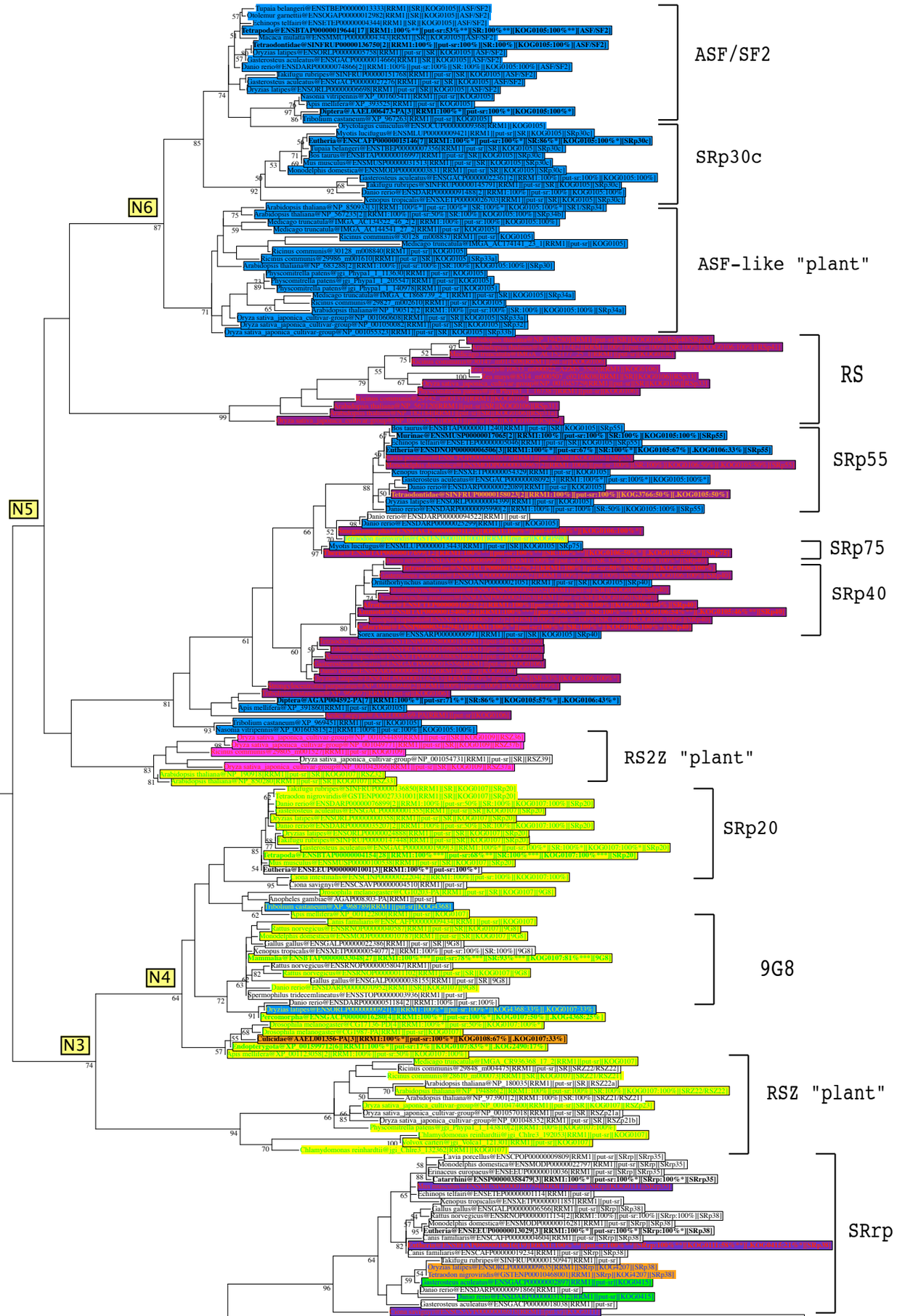
434x93



304x87



304x87



We computed a range of trees and identified ~12 subfamilies

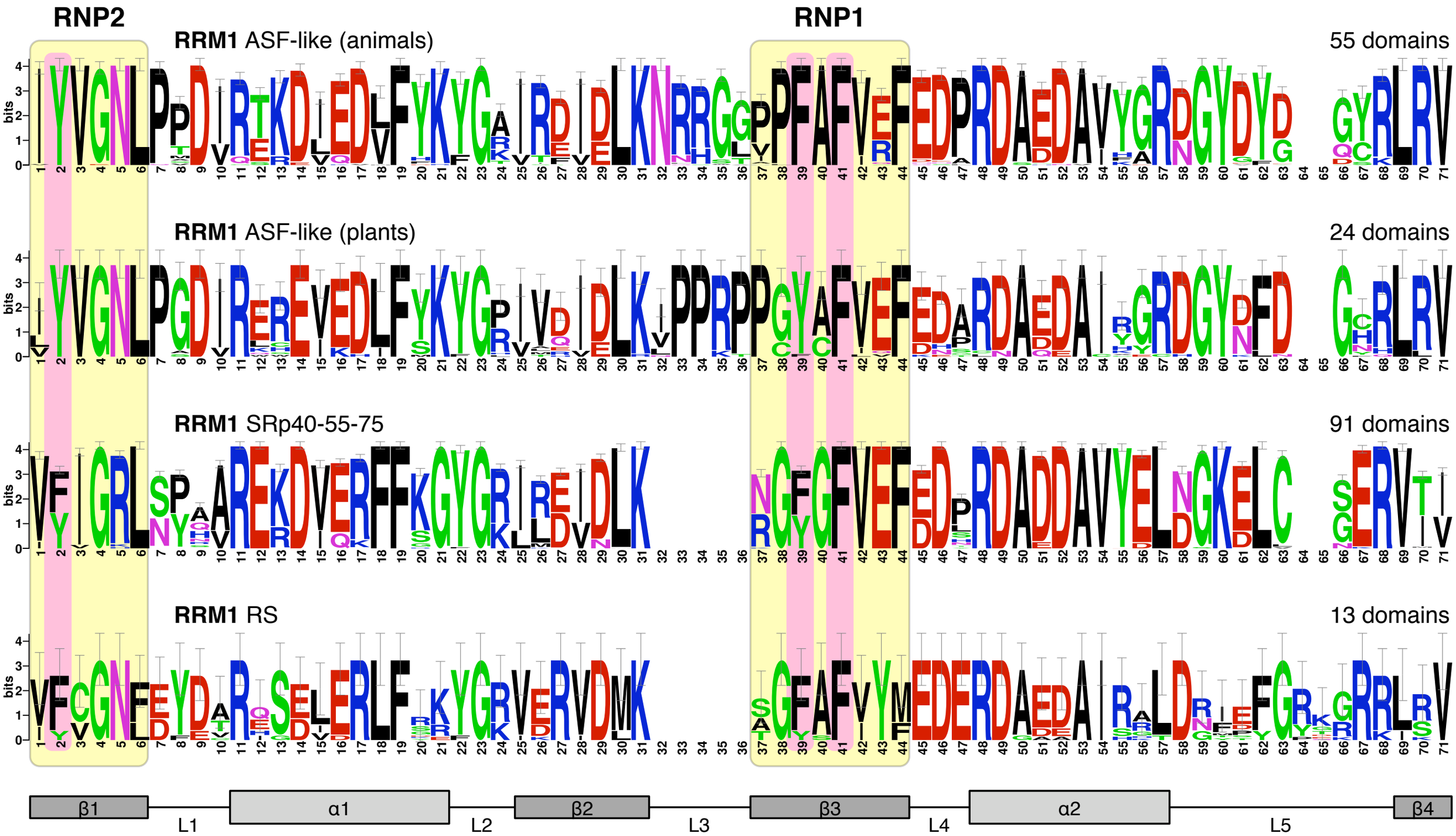
Data Set[a]	Sequence No. × Amino Acid No.	Heuristics	Model	Tree	Bootstrap Proportions for Nodes in Figure 3					
					N1	N2	N3	N4	N5	N6
Exhaustive	434 × 93	RAxML	WAG+Γ_4	Supplemental Figure S13	–[b]	<50	<50	<50[c]	<50[d]	<50
		TreeFinder	WAG+Γ_4	Supplemental Figure S14	–[b]	<50	<50	<50[c]	<50[d]	<50
Slow evolving	304 × 87	RAxML	WAG+Γ_4	Supplemental Figure S19	<50	68	74	64	<50[d]	87
		TreeFinder	WAG+Γ_4	Supplemental Figure S20	<50	62	85	84	55[d]	87
Slow-evolving−RS	292 × 85	RAxML	WAG+Γ_4	/[e]	<50	58	69	62	51[d]	84
		TreeFinder	WAG+Γ_4	/[e]	–[b]	64	69	88	<50	94
Slow-evolving−RS2Z	297 × 87	RAxML	WAG+Γ_4	/[e]	53	60	62	59	57	71
		TreeFinder	WAG+Γ_4	/[e]	<50	58	61	89	68	89
Slow-evolving−RS-RS2Z	285 × 85	RAxML	WAG+Γ_4	/[e]	<50	53	56	63	59	85
		TreeFinder	WAG+Γ_4	/[e]	–[b]	62	67	83	58	97
Inventory, all	270 × 82	RAxML	WAG+Γ_4	/[e]	–[b]	<50	–[f]	<50[c]	<50[d]	<50
		TreeFinder	WAG+Γ_4	/[e]	58	<50	–[f]	–[f]	–[f]	<50
Inventory, no gap	270 × 64	RAxML	WAG+Γ_4	Supplemental Figure S26	<50	<50	<50	<50[c]	–[f]	<50[c]
		TreeFinder	WAG+Γ_4	Supplemental Figure S27	64	<50	<50	<50[c]	–[f]	62[c]
		RAxML	LG+F+Γ_4	Supplemental Figure S28	<50	<50	–[f]	<50[c]	<50[d]	<50[c]
		PhyloBayes	CAT+Γ_4	Supplemental Figure S29	92	–[f]	–[f]	–[f]	<50[d]	<50[c]

[a]Data sets are described in the text. The relationships are based on the analysis of RRM1 domains. [b]Paraphyletic group due to outgroup domains. [c]Except for a few fast-evolving domains. [d]Actually paraphyletic due to RS2Z domains. [e]Tree not shown. [f]Node not recovered in the tree (i.e. the group is polyphyletic).

Varying sequence and/or subfamily sampling improved the statistical support for some nodes of interest.

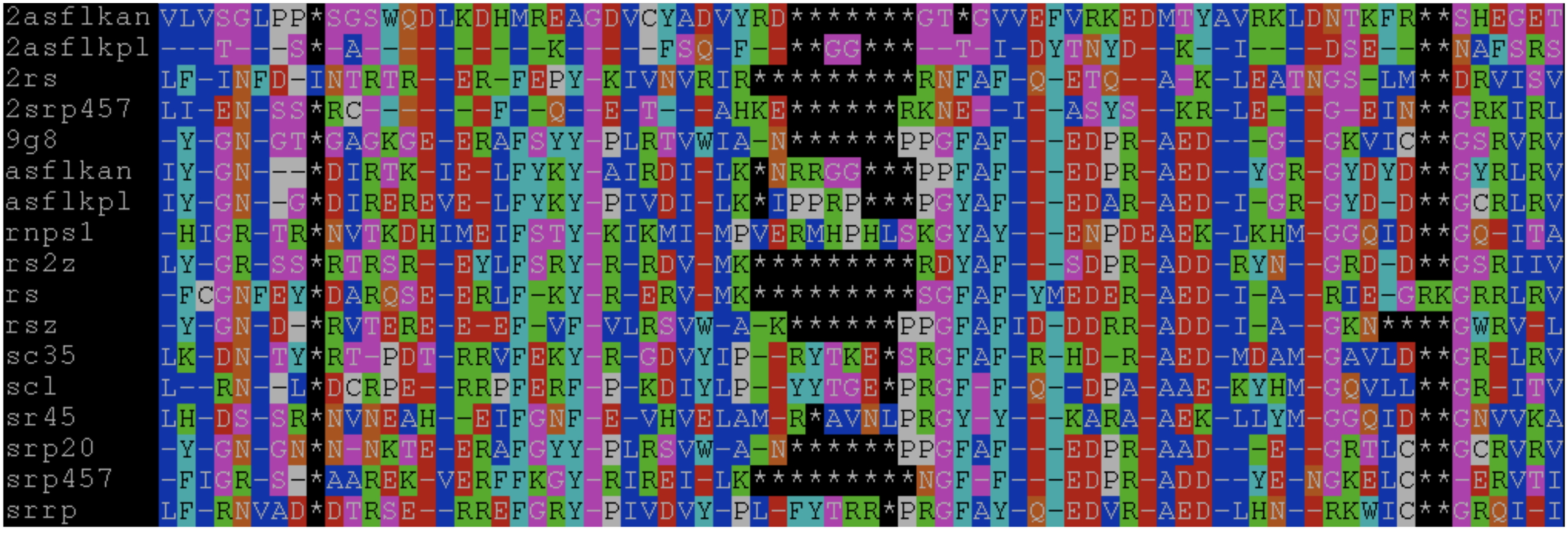
Changing heuristics sometimes affected the results.

Califice et al. (2012) *Plant Physiol* 158:546-560



For each subfamily, we computed a profile of the RRM on which the secondary structure was then mapped.

Califice et al. (2012) *Plant Physiol* 158:546-560



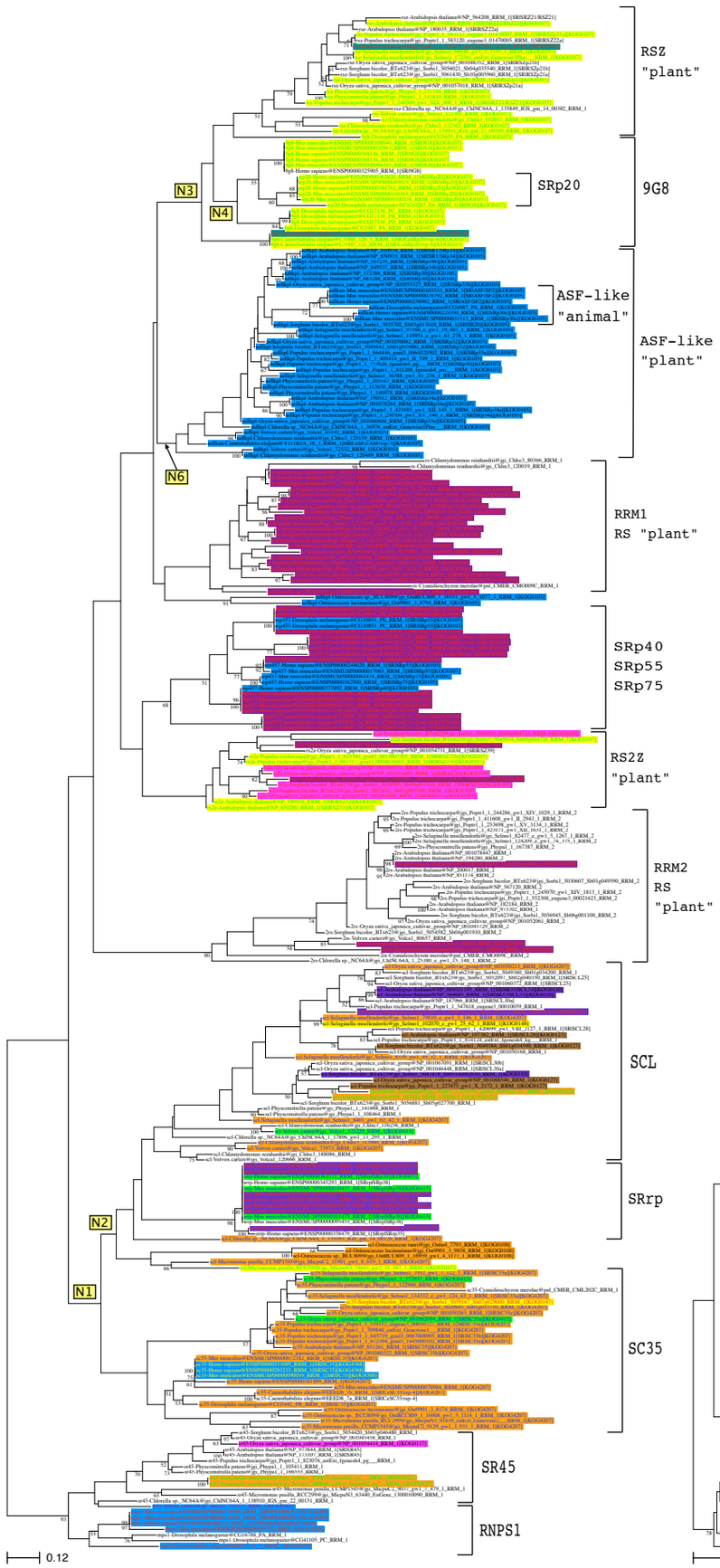
These profiles allowed us to build a structural alignment of the RRM domain across all subfamilies.

We also aligned the RRM2 of dual-RRM SR proteins, which helped us to deduce its evolutionary origin.

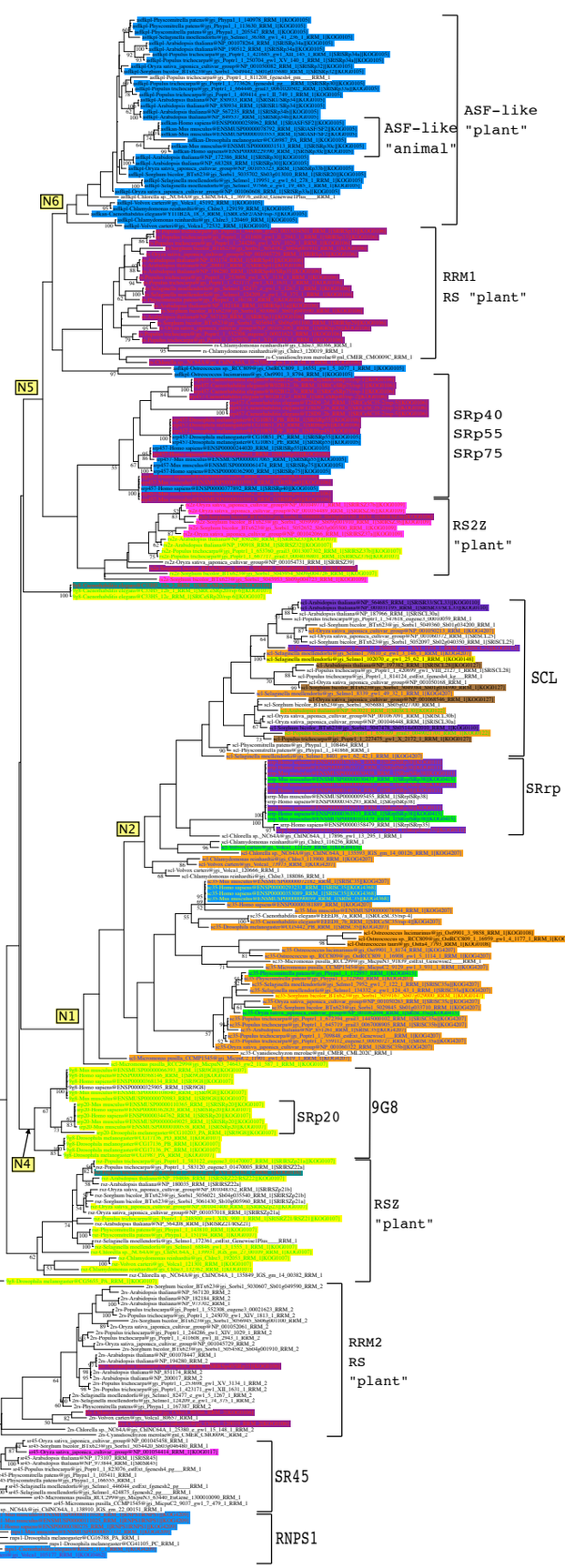
For validation, we used SR-optimized pHMMs to directly predict SR proteins in 20 model organisms.

270x64

RAxML/WAG

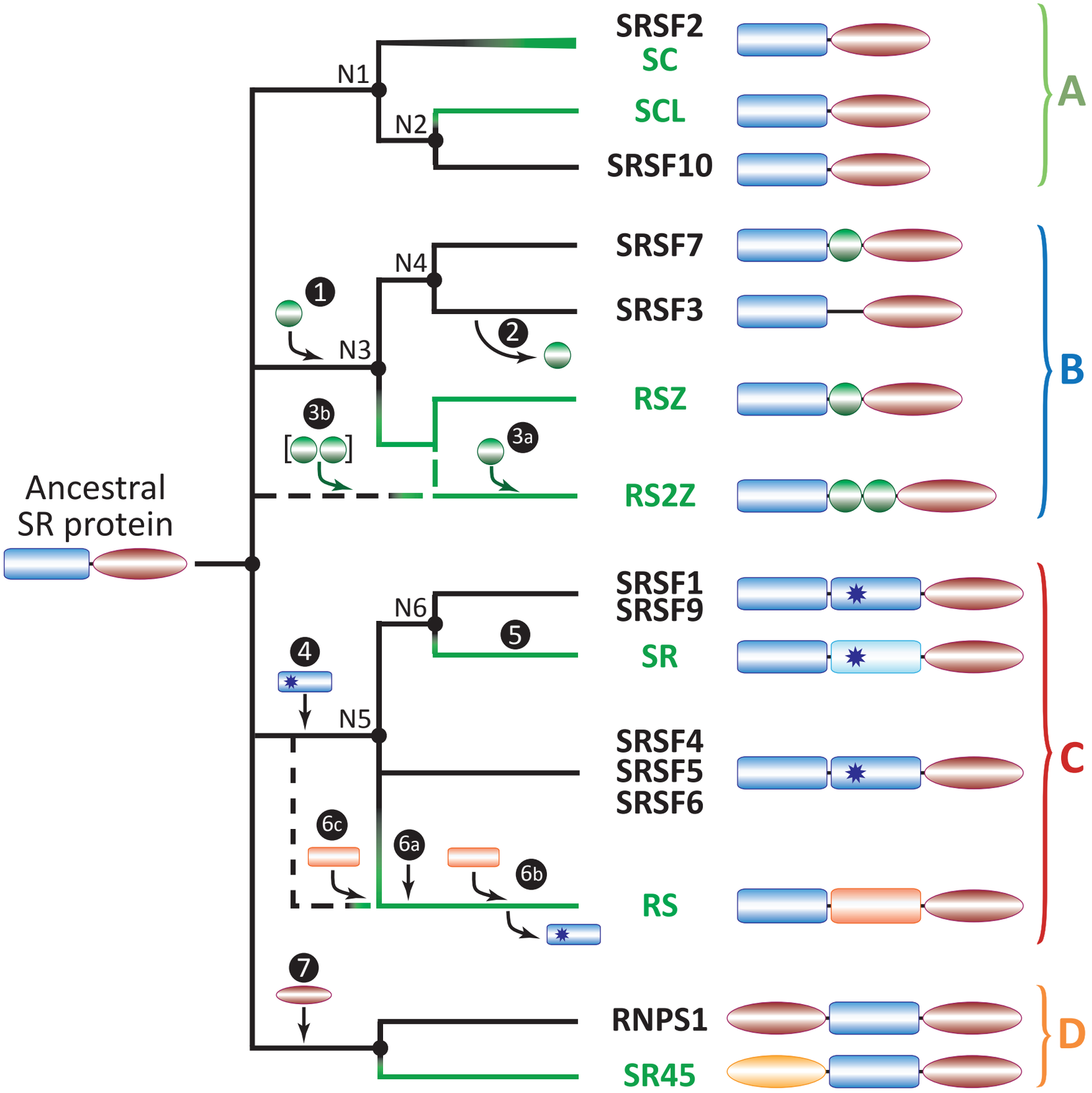


RAxML/LG



The corresponding RRM domains were structurally aligned.

However, the resulting trees were quite sensitive to the model, which confirms a lack of phylogenetic signal.



Nevertheless, we proposed a scenario for the diversification of SR proteins.

This hypothesis establishes new relationships between animal and plant SR proteins

Califice et al. (2012) *Plant Physiol* 158:546-560

Conclusions

1. SR splicing factors share a common origin and thus belong to a natural protein family.
2. They diversified early in eukaryotic evolution since the four principal architectures are shared between animals and plants.

Acknowledgments

Plant Physiol. 2012 Feb;158(2):546-60. Epub 2011 Dec 12.

A single ancient origin for prototypical serine/arginine-rich splicing factors.

Califice S, Baurain D, Hanikenne M, Motte P.

Laboratory of Functional Genomics and Plant Molecular Imaging and Centre for Assistance in Technology of Microscopy, Department of Life Sciences, Institute of Botany, University of Liège, B-4000 Liege, Belgium.

Abstract

Eukaryotic precursor mRNA splicing is a process involving a very complex RNA-protein edifice. Serine/arginine-rich (SR) proteins play essential roles in precursor mRNA constitutive and alternative splicing and have been suggested to be crucial in plant-specific forms of developmental regulation and environmental adaptation. Despite their functional importance, little is known about their origin and evolutionary history. SR splicing factors have a modular organization featuring at least one RNA recognition motif (RRM) domain and a carboxyl-terminal region enriched in serine/arginine dipeptides. To investigate the evolution of SR proteins, we infer phylogenies for more than 12,000 RRM domains representing more than 200 broadly sampled organisms. Our analyses reveal that the RRM domain is not restricted to eukaryotes and that all prototypical SR proteins share a single ancient origin, including the plant-specific SR45 protein. Based on these findings, we propose a scenario for their diversification into four natural families, each corresponding to a main SR architecture, and a dozen subfamilies, of which we profile both sequence conservation and composition. Finally, using operational criteria for computational discovery and classification, we catalog SR proteins in 20 model organisms, with a focus on green algae and land plants. Altogether, our study confirms the homogeneity and antiquity of SR splicing factors while establishing robust phylogenetic relationships between animal and plant proteins, which should enable functional analyses of lesser characterized SR family members, especially in green plants.

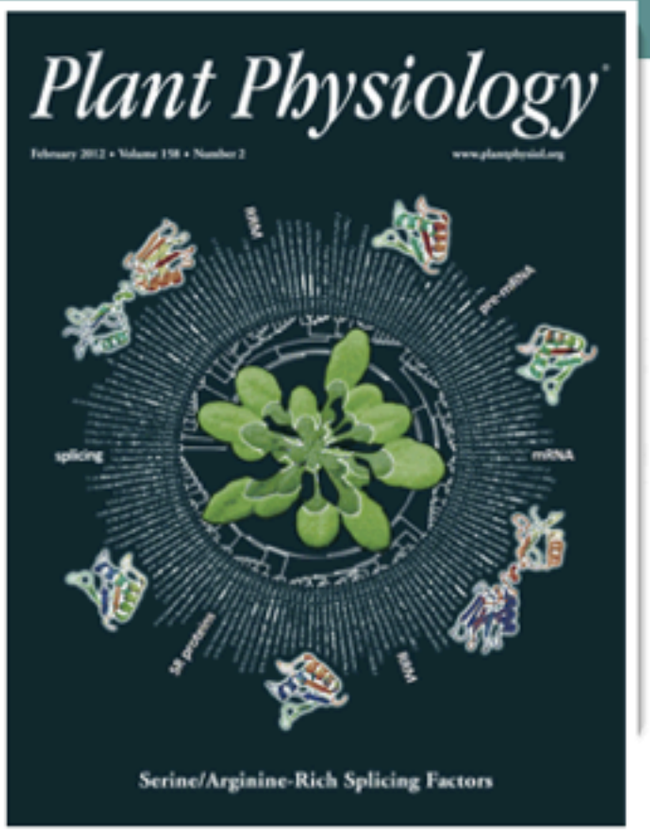
Plant Physiology®

AMERICAN SOCIETY OF PLANT BIOLOGISTS

Five-Year Impact Factor: 7.016
Most highly cited plant journal 2010: 55,626

User Name
User Name
Password
LOG-IN

Quick Search [advanced]
Author: (e.g. Smith, JS)
Keyword(s):
Year: Vol: Page: Go



Plant Physiology ® is an international journal devoted to physiology, biochemistry, cellular and molecular biology, genetics, biophysics, and environmental biology of plants.

Current Issue: February 2012
***Plant Physiology* Preview**
Search for Articles
Select an Issue from the Archive
Focus Collections
Upcoming Focus Issues
Classics Collection
Annual Update Collection
Editorials, Updates, and Featured Articles
Most-Read Articles
Most-Cited Articles
Permissions
Recognizing Our Authors

ABOUT PLANT PHYSIOLOGY
EDITORIAL BOARD AND STAFF
INSTRUCTIONS FOR AUTHORS
SUBMIT A MANUSCRIPT
RECOMMEND TO LIBRARY
LIBRARIANS
E-MAIL ALERTS (FREE)
RSS FEEDS (FREE)
HELP
PLANT GENE REGISTER

Meet Mike Blatt, incoming Editor-In-Chief of Plant Physiology!

Top 100 Journals — 100 Most Influential Journals of the Last 100 Years in Medicine and Biology — DBIO

F1000 Faculty of 1000 · CrossCheck Powered by iThenticate · Teaching Tools in Plant Biology — ideas to grow on · research4life HINARI · AGORA · OARE

American Society of Plant Biologists

FFa 8

A single ancient origin for prototypical serine/arginine-rich splicing factors.

Califice S, Baurain D, Hanikenne M, Motte P.
Plant Physiol. 2012 Feb; 158(2):546-60

Kathleen Hall, Washington University School of Medicine, MO, USA. F1000 Cell Biology

8 MUST READ

29 Feb 2012 | Technical Advance, New Finding

This comprehensive, state-of-the-art phylogenetic analysis of RNA recognition motifs (RRMs) leads the authors to the conclusion that the RRMs of serine/arginine-rich (SR) proteins arose from a single common ancestor. This massive collation of proteomes and the extraction of RRMs is a model of what can be done today with protein phylogeny and, from the perspective of RNA-binding proteins, it is a storehouse of information on the relationships between the ubiquitous RRMs. As an example, their total yield of RRMs is 12,023 from a total of 8042 proteins. A total of 556 RRMs come from humans.

Along the way, they identify prokaryotic RRMs with consensus RNP1 and RNP2 motifs, predominantly from cyanobacteria, that they suggest support the ancient lineage of the RRM. Curiously, RRMs are virtually absent in alpha-proteobacteria. The discussion of the evolution of the SR proteins is a model of how other RRM lineages could be evaluated, which hopefully the authors will provide.

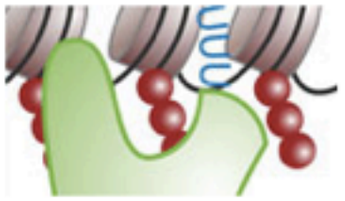
Competing Interests
None declared

Cite this evaluation

Cite this page

RELATED F1000 REPORTS

Controlling animal growth and body size – does fruit fly physiology point the way?
Savraj S. Grewal



The context of gene expression regulation
Johan H. Gibcus, Job Dekker

Lipids as conductors in